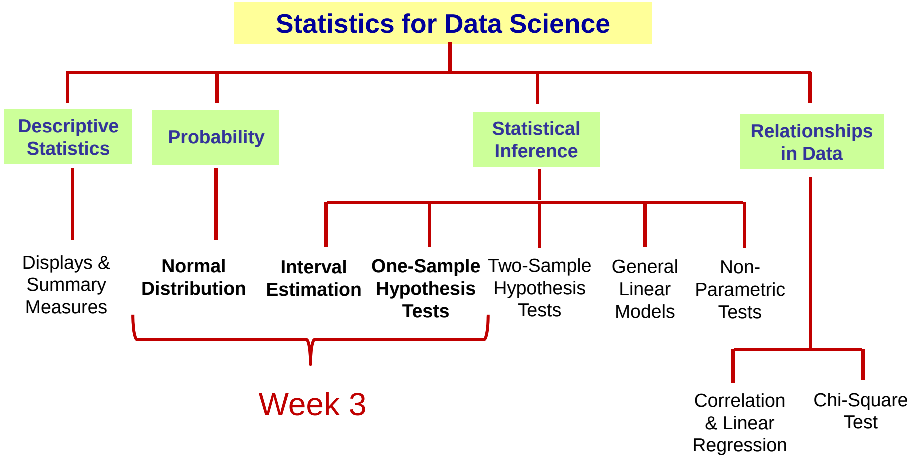
MATH 4044
Statistics for Data Science

Foundations of Inference

Course outline



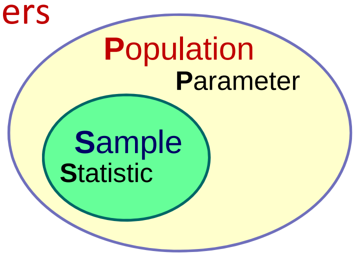
Field, A & Miles, J, *Discovering Statistics Using SAS*,
Chapter 2 (all except sections 2.6.4, 2.6.5) & Chapter 5 (all expect section 5.6)

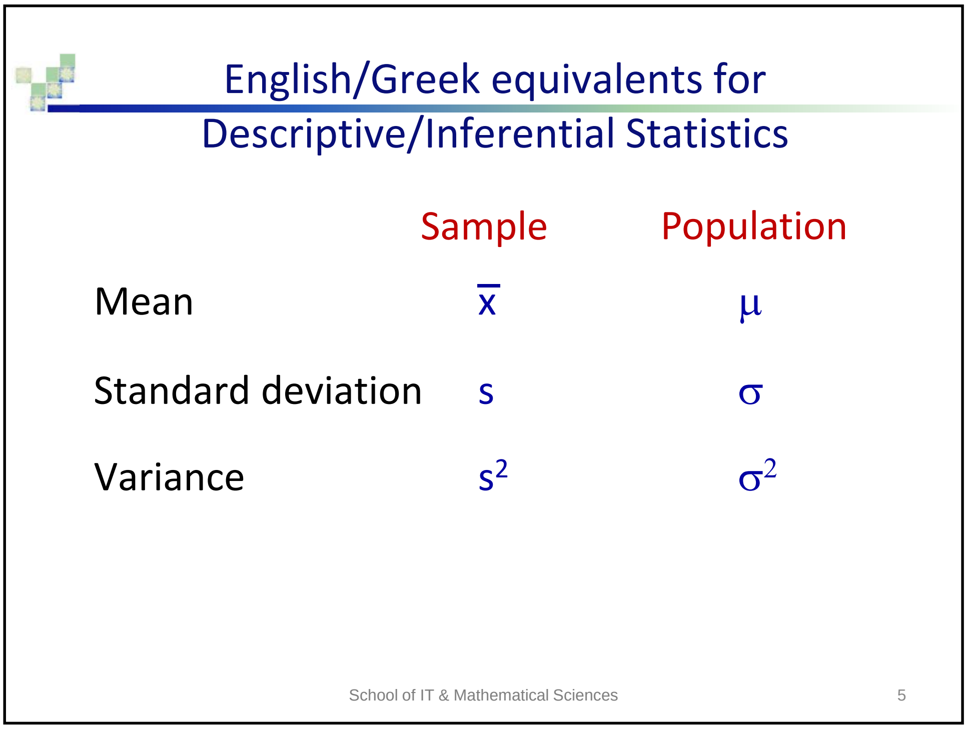
Topics to be covered

- Single-sample inference:
  - □ Sampling distributions
  - □ Central Limit Theorem (CLT)
  - □ Confidence intervals
  - □ Hypothesis tests
  - □ Checking conditions for inference

Statistical Inference

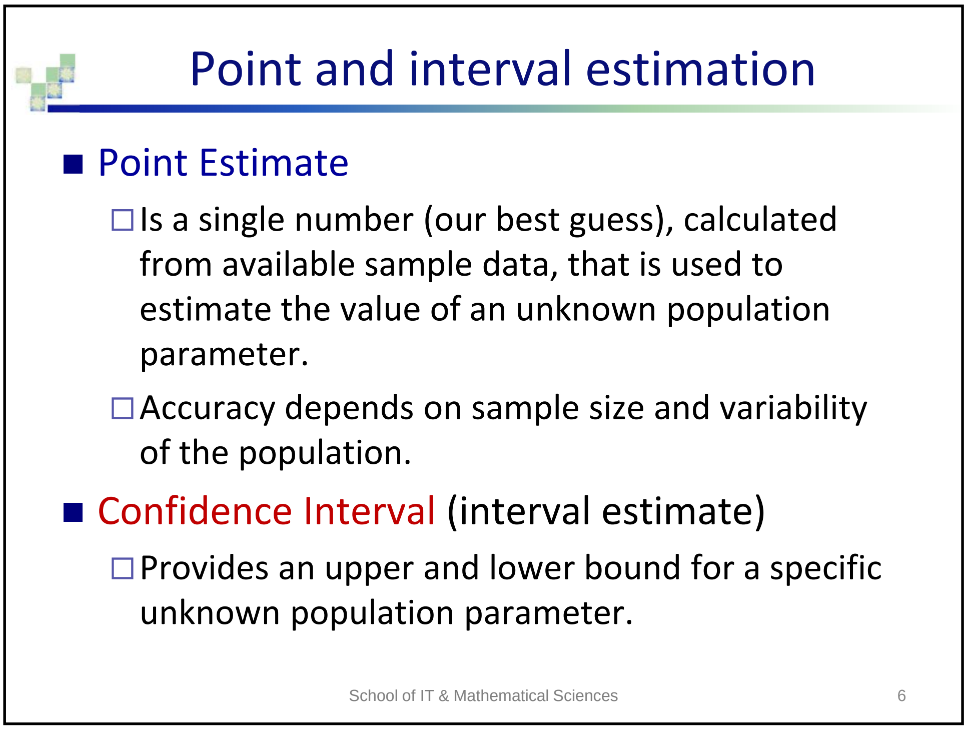
- A formal process that uses information from a sample to draw conclusions about a population.
- It also provides a statement of how much confidence can be placed in the conclusion.
- Conclusions about parameters are made using statistics.





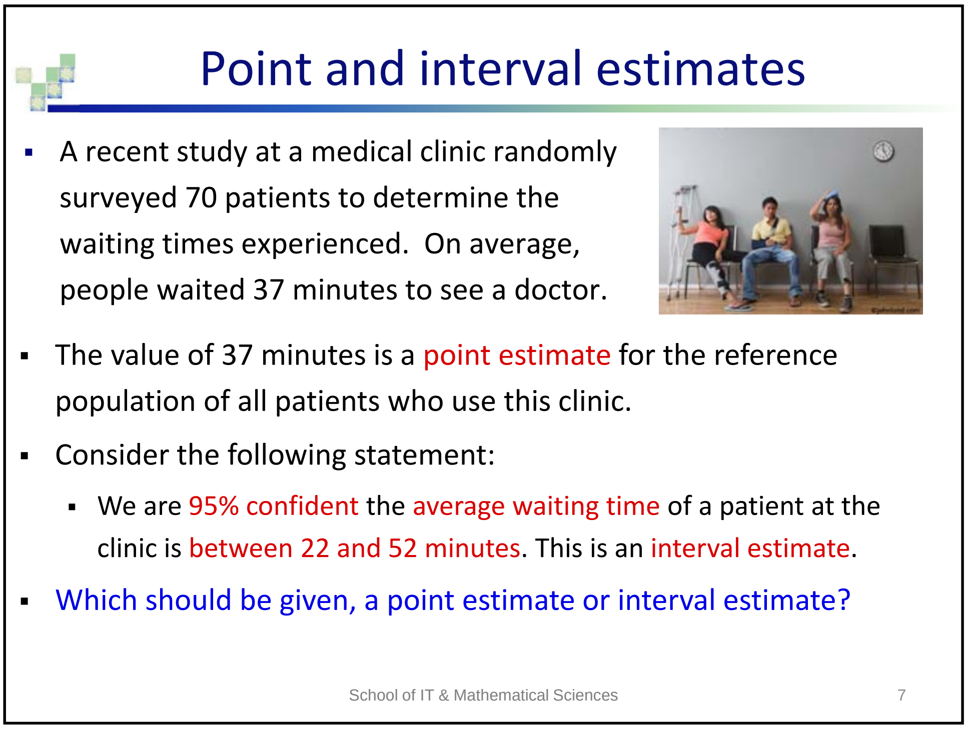
English/Greek equivalents for Descriptive/Inferential Statistics

	Sample	Population
Mean	$\bar{x}$	μ
Standard deviation	s	σ
Variance	s^2	σ^2



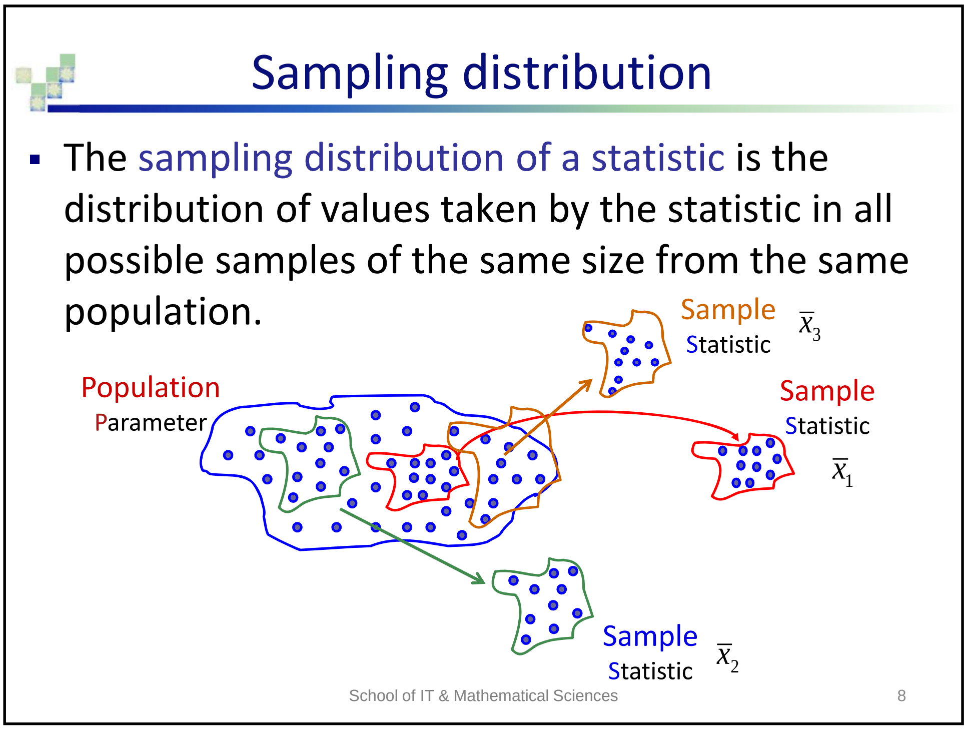
Point and interval estimation

- Point Estimate
  - □ Is a single number (our best guess), calculated from available sample data, that is used to estimate the value of an unknown population parameter.
  - □ Accuracy depends on sample size and variability of the population.
- Confidence Interval (interval estimate)
  - □ Provides an upper and lower bound for a specific unknown population parameter.



Point and interval estimates

- A recent study at a medical clinic randomly surveyed 70 patients to determine the waiting times experienced. On average, people waited 37 minutes to see a doctor.
- The value of 37 minutes is a point estimate for the reference population of all patients who use this clinic.
- Consider the following statement:
  - We are 95% confident the average waiting time of a patient at the clinic is between 22 and 52 minutes. This is an interval estimate.
- Which should be given, a point estimate or interval estimate?

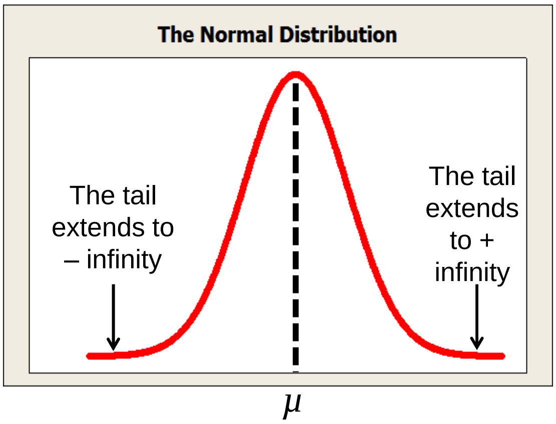


Sampling distribution

- The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

Normal Distribution

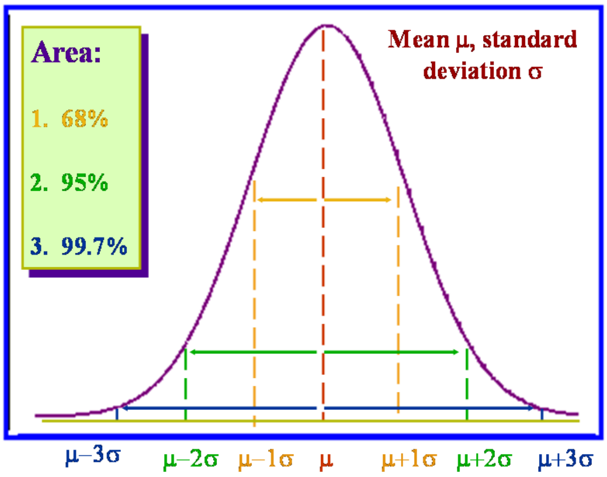
- A (large) *population* is said to have a *Normal distribution* when the frequencies of observations produce a histogram that follows the pattern of a smooth bell-shaped curve.

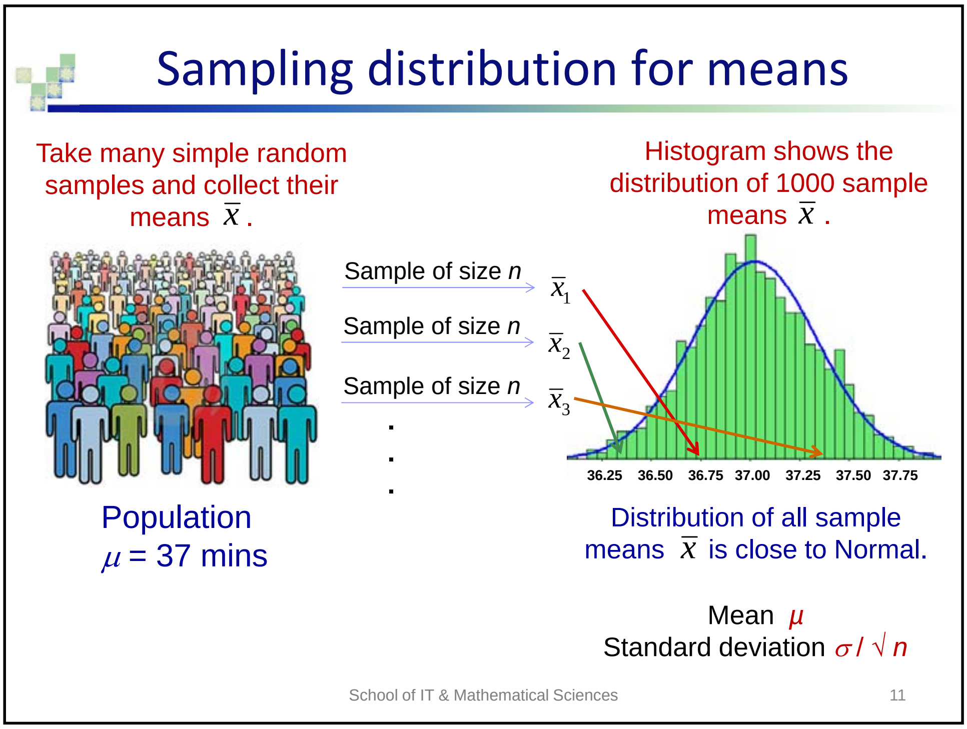

The Normal Distribution
The tail extends to – infinity
The tail extends to + infinity
μ

Many real-life data sets take this shape, hence the name given to this curve is `Normal'.

Standard Normal Distribution: a Normal distribution with $\mu = 0$ and $\sigma = 1$.

68 - 95 - 99.7 Rule





Sampling distribution for means

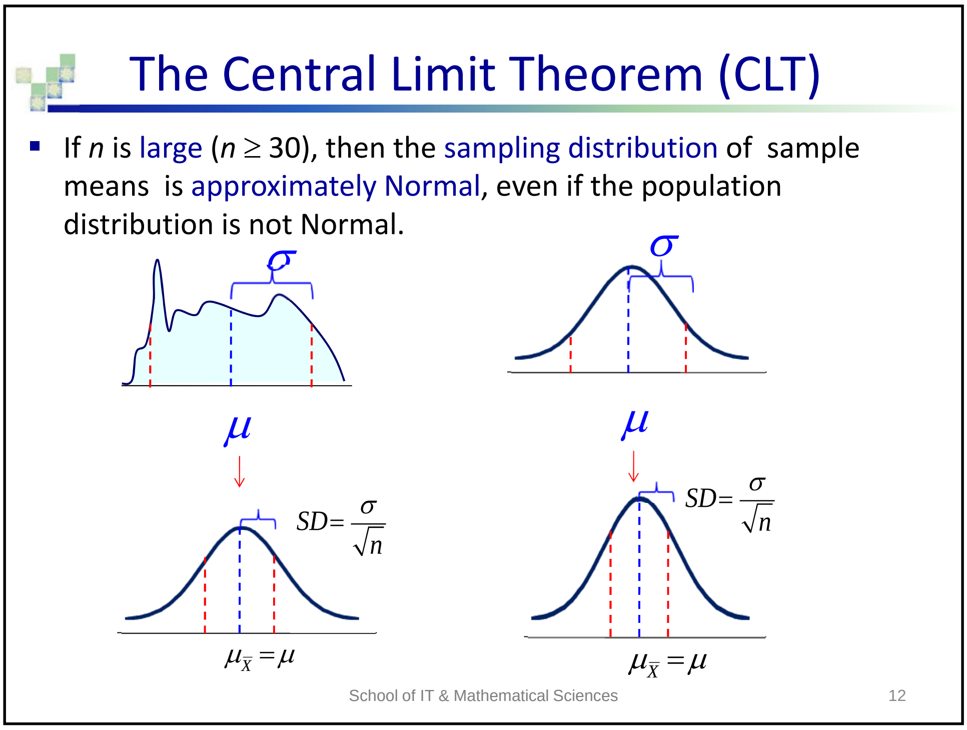
Take many simple random samples and collect their means $\overline{X}$.

Histogram shows the distribution of 1000 sample means $\overline{X}$.

Population
μ = 37 mins

Distribution of all sample means $\overline{X}$ is close to Normal.

Mean μ
Standard deviation $\sigma / \sqrt{n}$

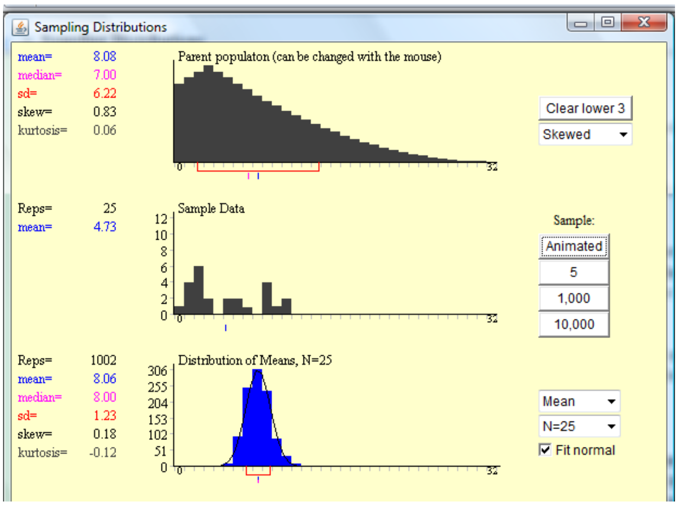


The Central Limit Theorem (CLT)

- If n is large ($n \geq 30$), then the sampling distribution of sample means is approximately Normal, even if the population distribution is not Normal.

A CLT simulation experiment

http://onlinestatbook.com/stat_sim/sampling_dist/index.html



Sampling Distributions

```
mean=      8.08
median=    7.00
sd=        6.22
skew=      0.83
kurtosis=  0.06

Reps=      25
mean=      4.73

Reps=      1002
mean=      8.06
median=    8.00
sd=        1.23
skew=      0.18
kurtosis=  -0.12
```

Sample: Animated

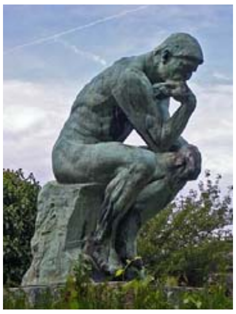
Skewed

Mean

N=25

☑ Fit normal

Thinking about sample means



- Means of random samples are less variable than individual observations.
- Means of random samples are more Normal than individual observations.

Standard Deviation vs Standard Error of the mean

- A measure of reliability or precision of $\overline{X}$ as an estimate of the population mean μ.
- The standard deviation of the sampling distribution of $\overline{X}$ is

$$SD = \frac{\sigma}{\sqrt{n}}$$

- When we estimate σ with *s*, we obtain the standard error of the mean:

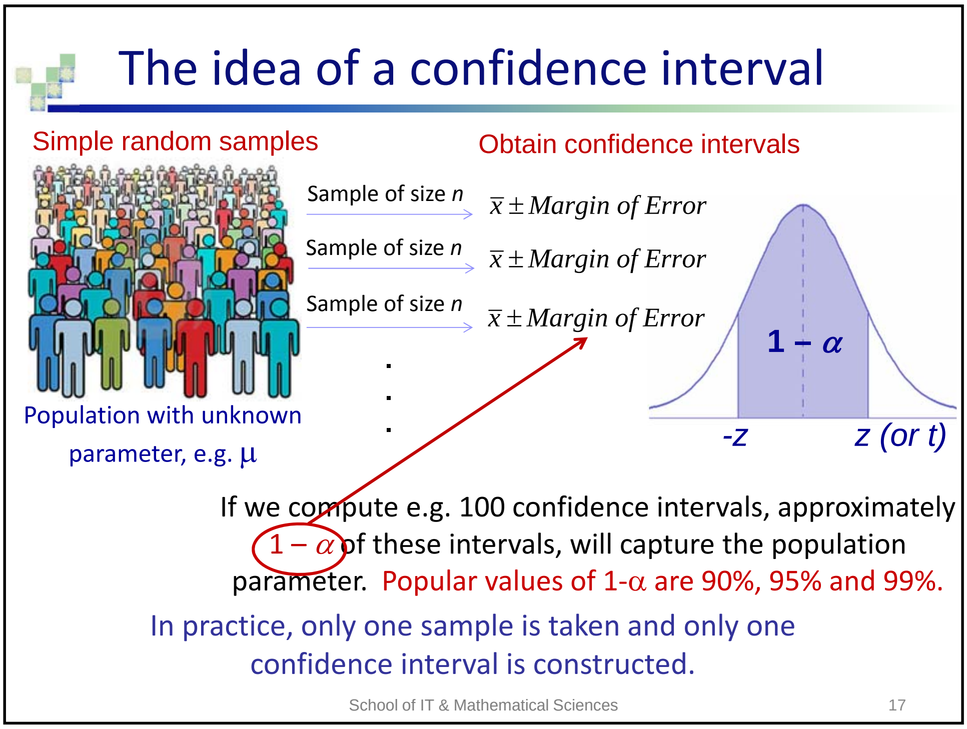
$$SE = \frac{s}{\sqrt{n}}$$

- We use either SD or SE to construct Confidence Intervals.

Confidence Interval (CI)

- Calculated from data, it is usually of the form

  Estimate ± margin of error
- The estimate is a sample statistic and the margin of error represents the accuracy of our guess for the parameter.
- A confidence level 1 – α gives the probability that the interval will capture the true parameter value in repeated samples.
  - □ It is the success rate of the method that produces the interval.
  - □ Chosen by the researcher, usually 90%, 95% or 99%.
- When we say we are 95% confident, we mean this:
  - □ We used a method that gives correct results 95% of the time.



The idea of a confidence interval

Simple random samples

Obtain confidence intervals

Sample of size *n* → $\bar{x} \pm Margin\ of\ Error$

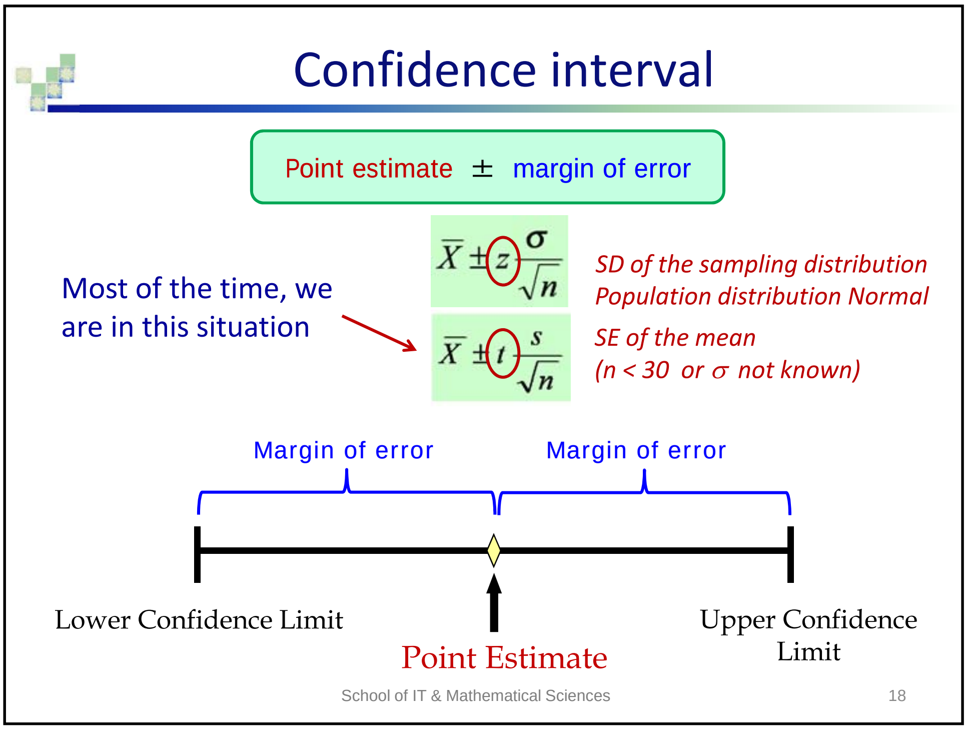
Sample of size *n* → $\bar{x} \pm Margin\ of\ Error$

Sample of size *n* → $\bar{x} \pm Margin\ of\ Error$

Population with unknown parameter, e.g. μ

If we compute e.g. 100 confidence intervals, approximately $1 - \alpha$ of these intervals, will capture the population parameter. Popular values of 1-α are 90%, 95% and 99%.

In practice, only one sample is taken and only one confidence interval is constructed.



Confidence interval

Point estimate ± margin of error

Most of the time, we are in this situation

$$\bar{X} \pm z\frac{\sigma}{\sqrt{n}}$$

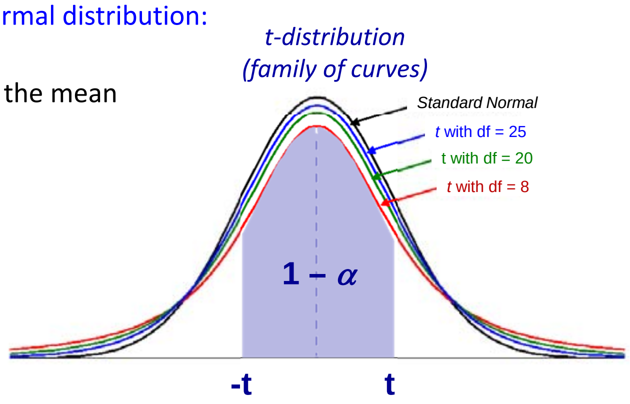
SD of the sampling distribution
Population distribution Normal

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

SE of the mean
(n < 30 or σ not known)

The t-distribution

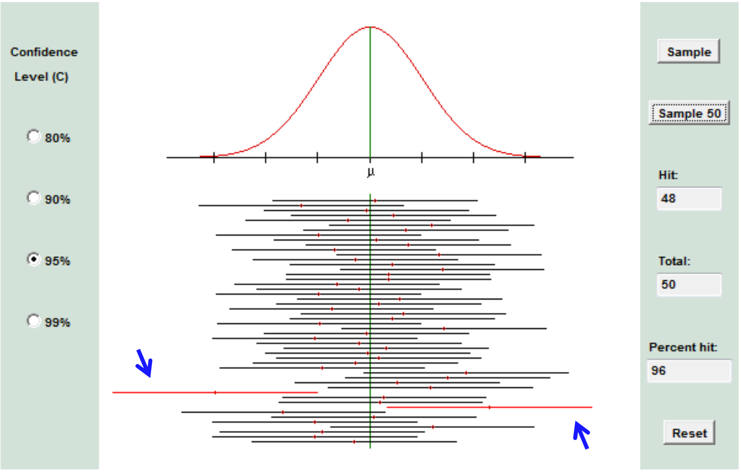
- Similarities to the Normal distribution:
  - Bell-shaped
  - Symmetric about the mean



- Dissimilarities:
  - Variance is larger than 1
  - Is a family of curves (depend on *degrees of freedom* = n-1)
  - As n becomes large, the t-distribution will look like the standard Normal.

A CI simulation experiment

http://bcs.whfreeman.com/ips4e/cat_010/applets/confidenceinterval.html


Confidence Level (C)

○ 80%
○ 90%
● 95%
○ 99%

Hit: 48

Total: 50

Percent hit: 96

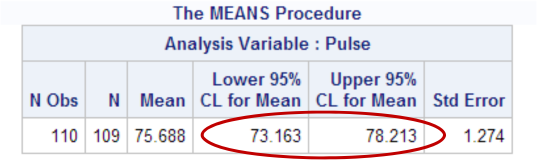
The intervals shown in red failed to capture the population mean μ

Requirements for inference about μ

- The sample is a simple random sample.
- Population is Normally distributed, or the sample size $n > 30$.
  - □ Check Normality using sample data and a P-P or Q-Q plot.

Example: Pulse rates

- What is the average pulse rate, in beats per minute?

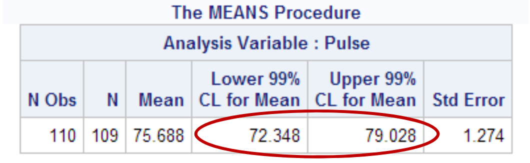


The MEANS Procedure

Analysis Variable : Pulse

N Obs	N	Mean	Lower 95% CL for Mean	Upper 95% CL for Mean	Std Error
110	109	75.688	73.163	78.213	1.274

- We are 95% confident that the *population* mean pulse rate is between 73.2 and 78.2 bpm.
- What if we change the confidence level to 99%?



The MEANS Procedure

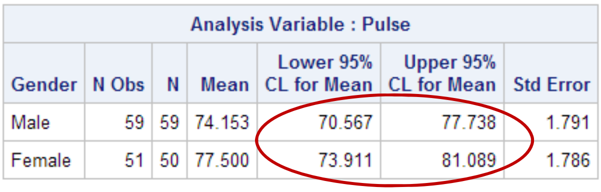
Analysis Variable : Pulse

N Obs	N	Mean	Lower 99% CL for Mean	Upper 99% CL for Mean	Std Error
110	109	75.688	72.348	79.028	1.274

The confidence interval becomes wider

Example: Pulse rates by *Gender*

- What is the average rate pulse rate for males? What is it for females?



Analysis Variable : Pulse						
Gender	N Obs	N	Mean	Lower 95% CL for Mean	Upper 95% CL for Mean	Std Error
Male	59	59	74.153	70.567	77.738	1.791
Female	51	50	77.500	73.911	81.089	1.786

- We are 95% confident that the *population* mean pulse is:
  - □ Between 70.6 and 77.7 bpm for males;
  - □ Between 73.9 and 81.1 bpm for females.
- Based on our result, does the pulse rate appear to depend on gender?

Confidence intervals using SAS

Using PROC MEANS:

```
proc format;
    value gender 1='Male' 2='Female';
run;

proc means data=mydata.pulse_rates
n mean clm stderr maxdec=3 printalltypes alpha=0.01;
    format Gender gender.;
    var Pulse;
    class Gender;
run;
```

Assigning new labels

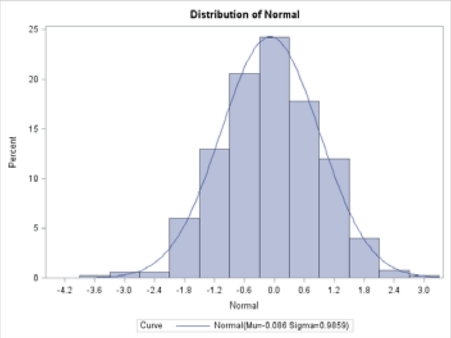
To get the grand mean as well as means broken down by Gender in one listing

Specifying confidence level; if not specified, $\alpha = 0.05$ by default

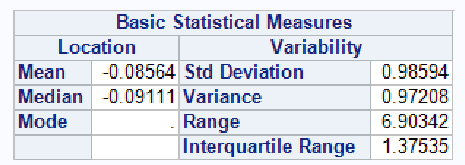
Checking Normality using a P-P or Q-Q plot

- A P-P plot shows ordered data against ‘ideal’ perfectly spaced Normal values (quantiles for a Q-Q plot).
- An approximate straight line is an indication of an approximate Normal distribution.
- Easy to use and to interpret.
- Non-Normality is concluded only for clear curved departures from a straight line fit.
- Formal tests are available (e.g. Kolmogorov-Smirnov, Shapiro-Wilks).

Example: Looking at the histogram

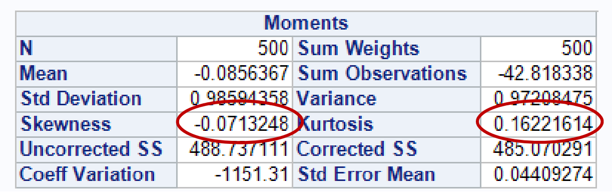


Tasks > Capability > Histograms…



Basic Statistical Measures			
Location		Variability	
Mean	-0.08564	Std Deviation	0.98594
Median	-0.09111	Variance	0.97208
Mode	.	Range	6.90342
		Interquartile Range	1.37535

This data was sampled from a standard Normal distribution.

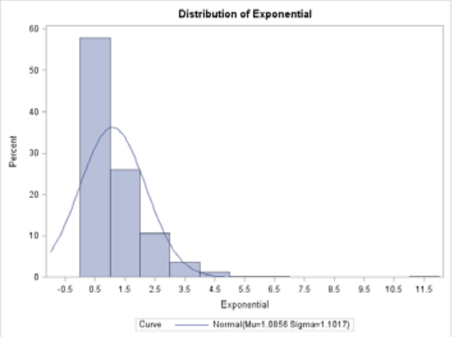


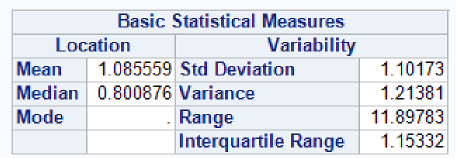
Moments			
N	500	Sum Weights	500
Mean	-0.0856367	Sum Observations	-42.818338
Std Deviation	0.98594358	Variance	0.97208475
Skewness	-0.0713248	Kurtosis	0.16221614
Uncorrected SS	488.737111	Corrected SS	485.070291
Coeff Variation	-1151.31	Std Error Mean	0.04409274

For a Normal distribution:

Skewness = 0
Kurtosis = 0

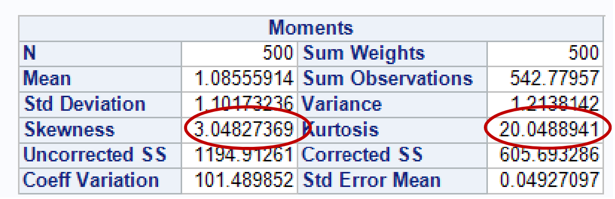
Example: Looking at the histogram





Basic Statistical Measures			
Location		**Variability**	
Mean	1.085559	Std Deviation	1.10173
Median	0.800876	Variance	1.21381
Mode	.	Range	11.89783
		Interquartile Range	1.15332

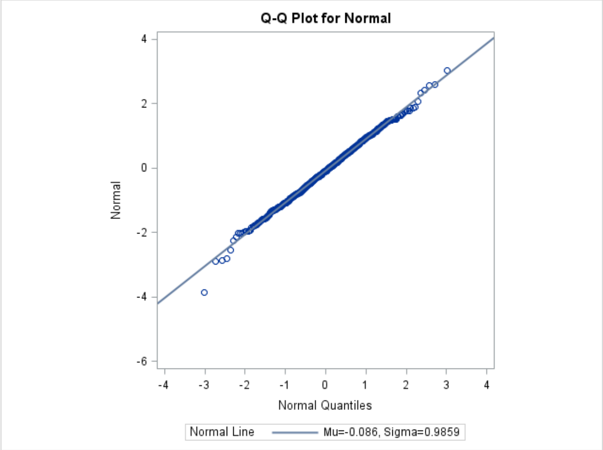
This data came from a distribution that is not Normal.



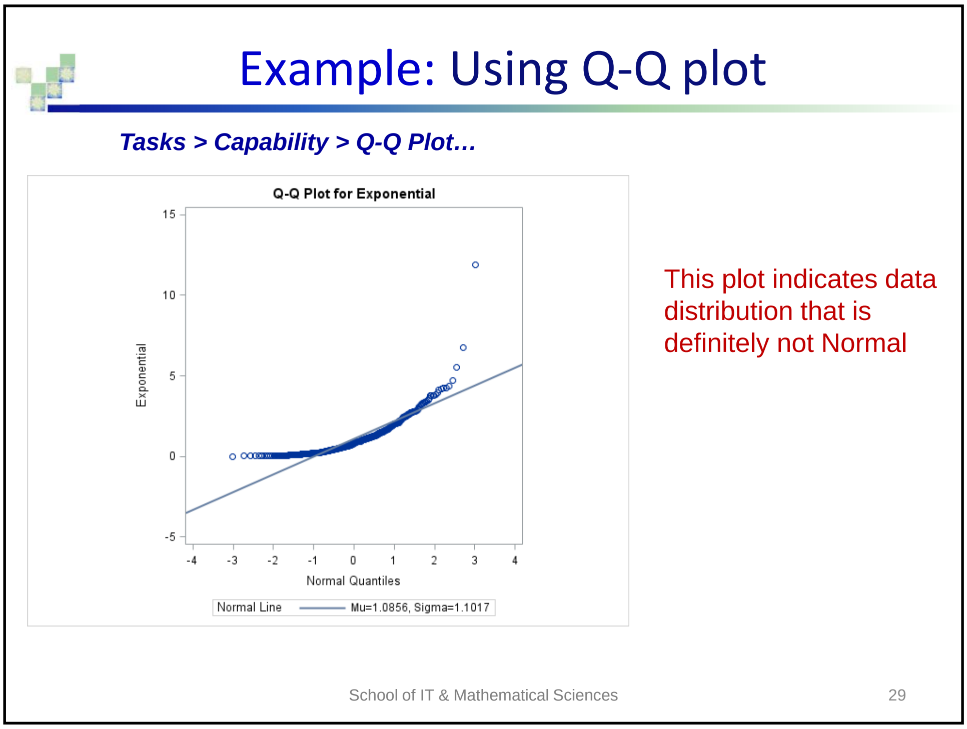
Moments			
N	500	Sum Weights	500
Mean	1.08555914	Sum Observations	542.77957
Std Deviation	1.10173236	Variance	1.2138142
Skewness	3.04827369	Kurtosis	20.0488941
Uncorrected SS	1194.91261	Corrected SS	605.693286
Coeff Variation	101.489852	Std Error Mean	0.04927097

Example: Using a Q-Q plot

Tasks > Capability > Probability Plots…



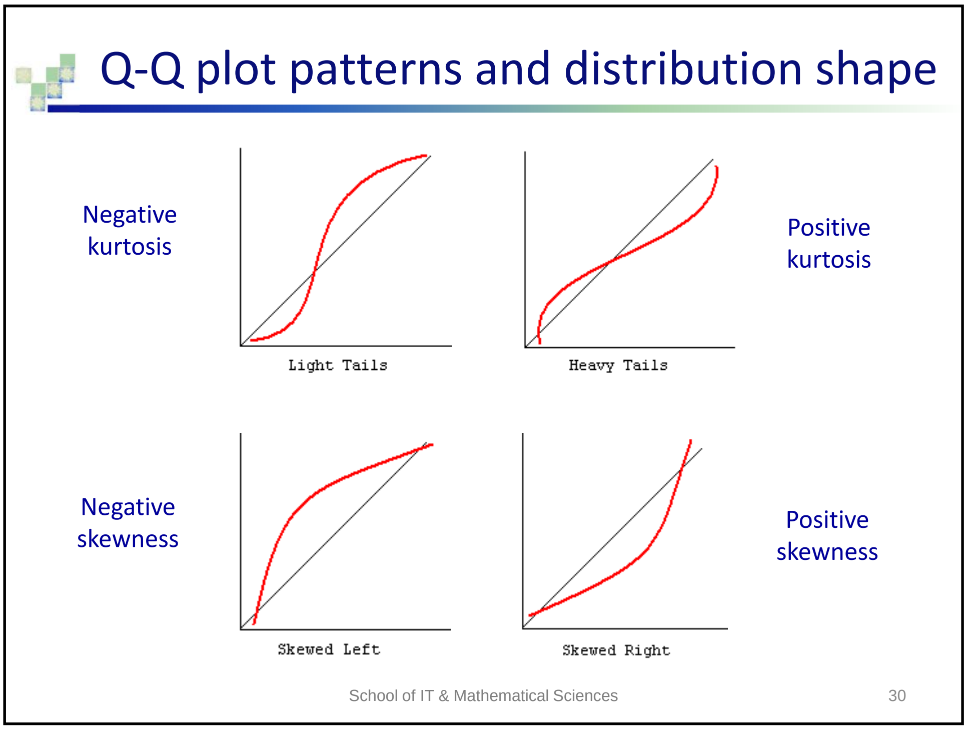
This plot indicates an approximately Normal data distribution



Example: Using Q-Q plot

Tasks > Capability > Q-Q Plot…

This plot indicates data distribution that is definitely not Normal



Q-Q plot patterns and distribution shape

Negative kurtosis

Positive kurtosis

Negative skewness

Positive skewness

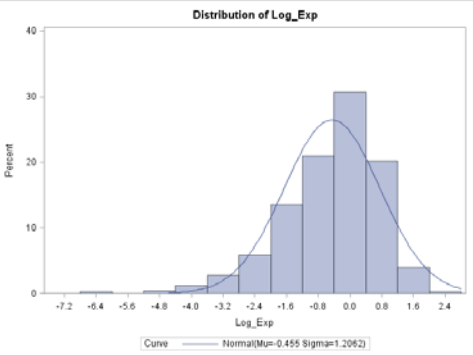
How can we make data ‘Normal’?

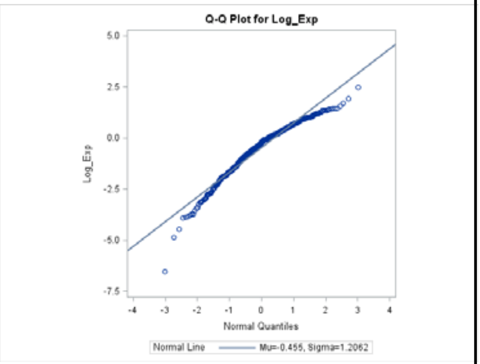
- If the distribution is skewed to the right, then one of the following transformations may be considered:

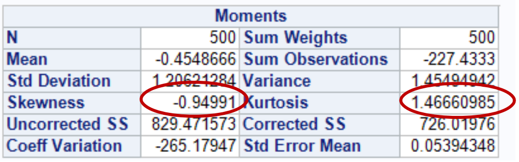
$$\sqrt{X} \qquad \log X \qquad 1/\sqrt{X} \qquad 1/X$$

- These transformation will pull in the long right tail and push out the short left tail, making the distribution more nearly symmetric.

Example: A log transformation







Moments			
N	500	Sum Weights	500
Mean	-0.4548666	Sum Observations	-227.4333
Std Deviation	1.20621284	Variance	1.45494942
Skewness	-0.94991	Kurtosis	1.46660985
Uncorrected SS	829.471573	Corrected SS	726.01976
Coeff Variation	-265.17947	Std Error Mean	0.05394348

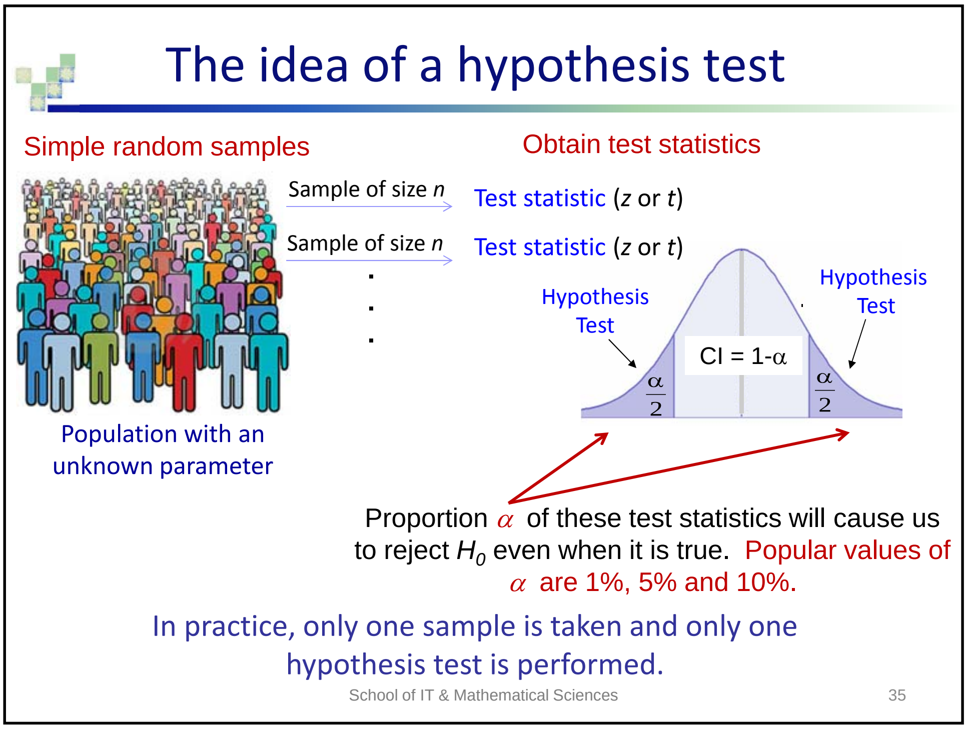
Transformed data has distribution much closer to Normal.

Brain Break

UNDER STRICT LABORATORY CONDITIONS, RESEARCH CONCLUDES THAT, IN SPITE OF BEING WATCHED, POTS DO EVENTUALLY BOIL.
search ID: cgon660
© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com

Hypothesis testing

- In statistics, a hypothesis is a claim or statement about a particular characteristic of a population.
  - E.g. A claim about a population parameter.
- A hypothesis test (or test of significance) is a procedure to test a claim about a population, e.g.
  - 5% of males suffer colour blindness.
  - Normal body temperature is 37 degrees Celsius.
  - Echinacea helps fight colds by boosting the immune system.
- Rare Event Rule:
  - If, under a given assumption, the probability of a particular observed event is small, we conclude the assumption may not be correct.



The idea of a hypothesis test

Simple random samples

Obtain test statistics

Sample of size *n* → Test statistic (*z* or *t*)

Sample of size *n* → Test statistic (*z* or *t*)

.
.
.

Population with an unknown parameter

Hypothesis Test

Hypothesis Test

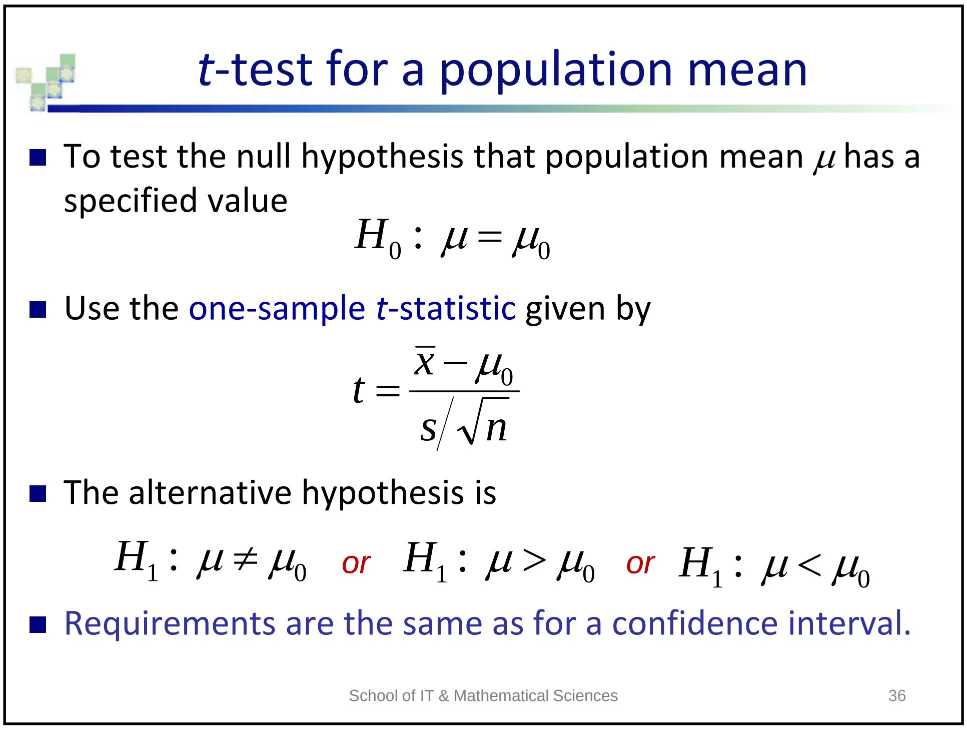
CI = 1-α

$\frac{\alpha}{2}$

$\frac{\alpha}{2}$

Proportion α of these test statistics will cause us to reject H_0 even when it is true. Popular values of α are 1%, 5% and 10%.

In practice, only one sample is taken and only one hypothesis test is performed.



t-test for a population mean

- To test the null hypothesis that population mean μ has a specified value

$$H_0 : \mu = \mu_0$$

- Use the one-sample *t*-statistic given by

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

- The alternative hypothesis is

$$H_1 : \mu \neq \mu_0 \quad \text{or} \quad H_1 : \mu > \mu_0 \quad \text{or} \quad H_1 : \mu < \mu_0$$

- Requirements are the same as for a confidence interval.

Test statistic and *P*-value

- A test statistic calculated from sample data measures how far the data diverge from the null hypothesis H_0.
  - □ Large values of the statistic show that the data are far from what we would expect if H_0 were true.

$$\text{test statistic} = \frac{\text{variance explained by the model}}{\text{variance not explained by the model}} = \frac{\text{effect}}{\text{error}}$$

- *P*-value is the probability, computed assuming H_0 is true, that the test statistic would take a value as extreme as or more extreme than that actually observed.
  - □ The smaller the *P*-value, the stronger the evidence against H_0 provided by the data.

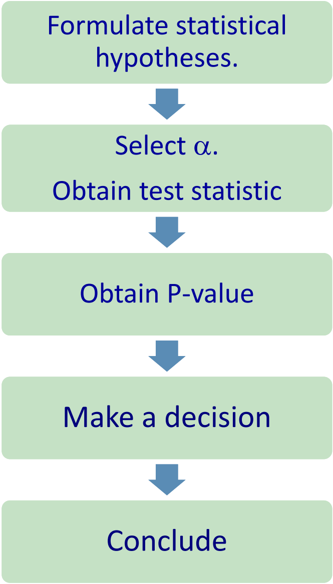
Errors in hypothesis tests

There are four possible scenarios in a hypothesis test:

		Test Conclusion	
Truth		Do not reject H_0	Reject H_0 in favour of H_1
	H_0 true	OK	Type I Error
	H_1 true	Type II Error	OK

- □ Type I errors occur when you reject H_0 as being false when H_0 is really true.
- □ Type II errors occur when you ~~accept~~ fail to reject H_0 as being true when H_0 is really false.
- □ Hypothesis tests are designed so as to reduce the chances of making Type I errors.

Steps in statistical hypothesis testing



$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$

Null hypothesis

Alternative hypothesis

$$\alpha = 0.05, \quad t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

α = 0.10, 0.05 or 0.01

Likelihood of a result at least as extreme, assuming H_0 is true

Small P-value implies there is an effect

Reject H_0 if P-value < α
Fail to Reject H_0 otherwise

α defines “rare” events (evidence against H_0)

Interpret in terms of research question

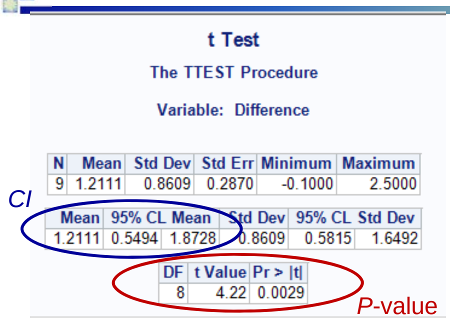
Example: Is the drug effective?

- The table below shows the difference in weight loss (kg) for nine subjects when taking the drug mCPP compared to taking a placebo.
- Does the drug mCPP affect weigh loss?

Subject	1	2	3	4	5	6	7	8	9
Difference	1.2	1.6	0.4	1.4	2.1	0.3	-0.1	2.5	1.5

APPARENTLY YOU IGNORED MY ADVICE AND GOT NO EXERCISE.
© 1990 United Feature Syndicate, Inc.
BUT YOU'RE IN PERFECT HEALTH, WHICH REALLY ANNOYS ME PROFESSIONALLY.
I'M PRESCRIBING TWO PACKS OF CIGARETTES PER DAY... DON'T CROSS ME AGAIN.

Example: Is the drug effective?



t Test

The TTEST Procedure

Variable: Difference

N	Mean	Std Dev	Std Err	Minimum	Maximum
9	1.2111	0.8609	0.2870	-0.1000	2.5000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1.2111	0.5494	1.8728	0.8609	0.5815	1.6492

DF	t Value	Pr > |t|
8	4.22	0.0029

CI

P-value

Tasks > ANOVA > T-test...
[This task uses PROC TTEST]

Hypotheses being tested:

$$H_0 : \mu = 0$$
$$H_1 : \mu \neq 0$$
$$\alpha = 0.05$$

Since the *P*-value = 0.0029 < 0.05, H_0 is rejected.

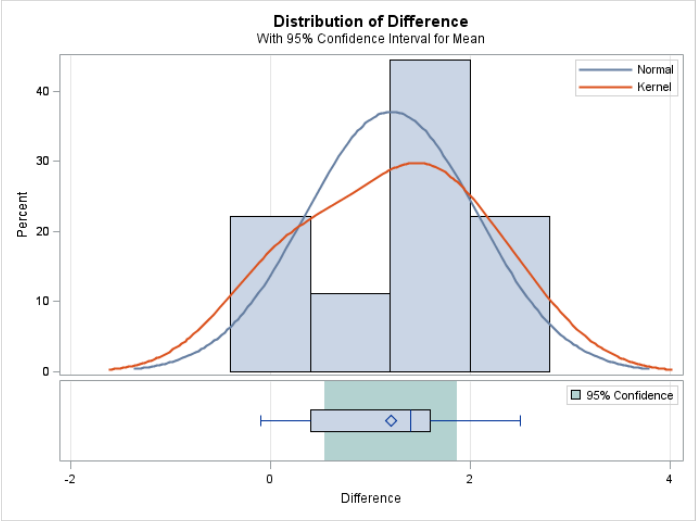
At 5% significance level, there is enough statistical evidence to conclude that the mean weight loss difference is not zero.

The result is significant at 5% level.

From the 95% confidence interval (0.5494, 1.8728), the mean difference in weight loss is actually positive; the drug is effective.

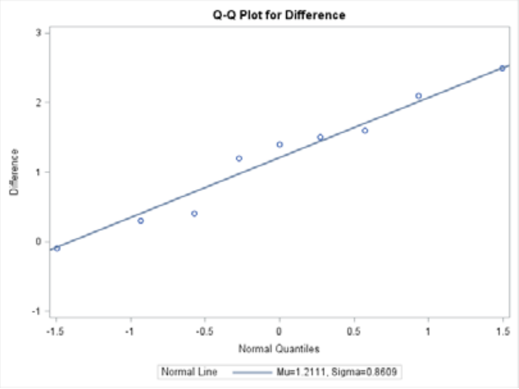
Example: Is the drug effective?

Tasks > ANOVA > T-test [This task uses PROC TTEST]

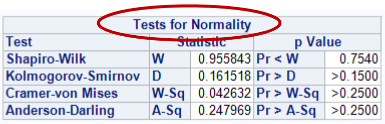


Example: Is the drug effective?

Tasks > Describe > Distribution Analysis… [This task uses PROC UNIVARIATE]



The plot shows an approximately straight line pattern.



Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.955843	Pr < W	0.7540
Kolmogorov-Smirnov	D	0.161518	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.042632	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.247969	Pr > A-Sq	>0.2500

H_0: Data comes from a Normal distribution

H_1: Data does not come from a Normal distribution

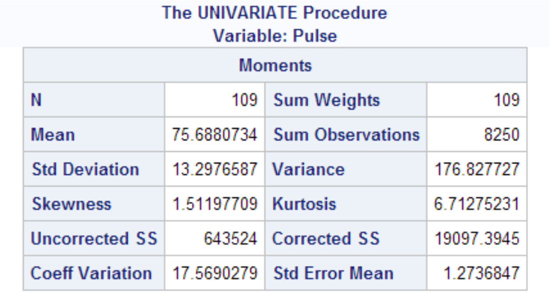
$\alpha = 0.05$

Since P-value > 0.05, H_0 can't be rejected.

We can assume Normality.

It made sense to proceed with a one-sample t-test. ✓

Example: Pulse rates



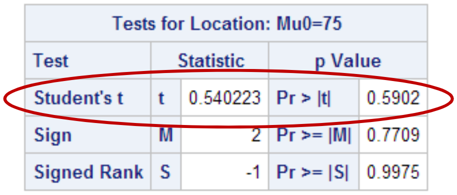
The UNIVARIATE Procedure

Variable: Pulse

Moments			
N	109	Sum Weights	109
Mean	75.6880734	Sum Observations	8250
Std Deviation	13.2976587	Variance	176.827727
Skewness	1.51197709	Kurtosis	6.71275231
Uncorrected SS	643524	Corrected SS	19097.3945
Coeff Variation	17.5690279	Std Error Mean	1.2736847

Hypotheses being tested:

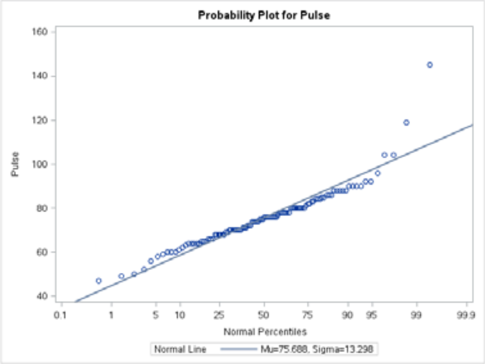
$$H_0: \mu = 75$$
$$H_1: \mu \neq 75$$
$$\alpha = 0.05$$



Tests for Location: Mu0=75				
Test	**Statistic**		**p Value**	
Student's t	t	0.540223	Pr > |t|	0.5902
Sign	M	2	Pr >= |M|	0.7709
Signed Rank	S	-1	Pr >= |S|	0.9975

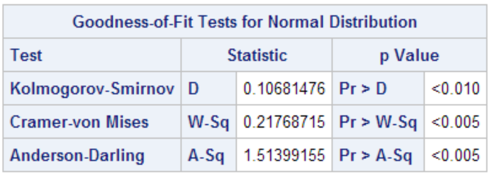
At 5% significance level, there is not enough statistical evidence to conclude that the population mean pulse rate for young adults is different from 75 bpm.

Example: Pulse rates



Based on our sample, the distribution of Pulse rates for the population of young adults can't be assumed to be Normal.

However, we have a large sample (*n* = 109) so the conditions for inference are satisfied.



Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.10681476	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.21768715	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	1.51399155	Pr > A-Sq	<0.005

It made sense to proceed with a one-sample t-test.

Example: Pulse rates

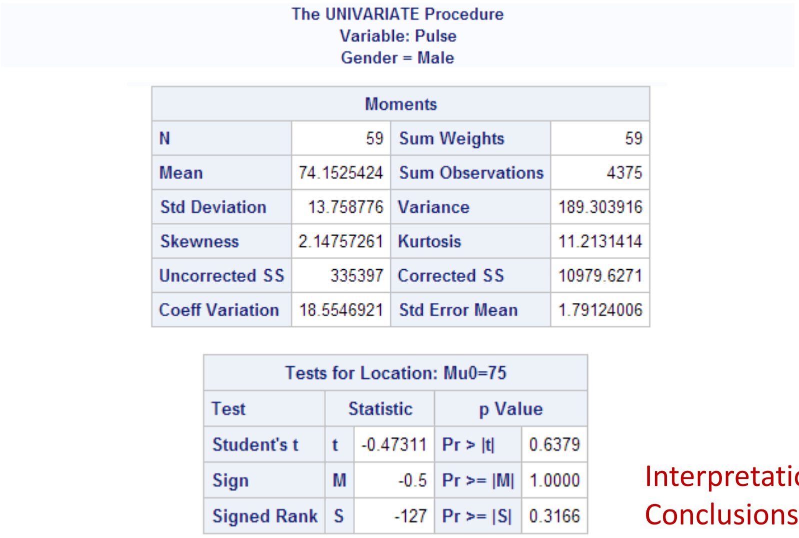
Using PROC UNIVARIATE:

Hypothesised mean value; Assumed to be zero if no value is specified

```
proc univariate data=mydata.pulse_rates mu0=75;
    var Pulse;
    histogram / normal;
    probplot / normal(mu=est sigma=est);
run;
```

Mean and standard deviation estimated from data

Exercise: Pulse rates by *Gender*



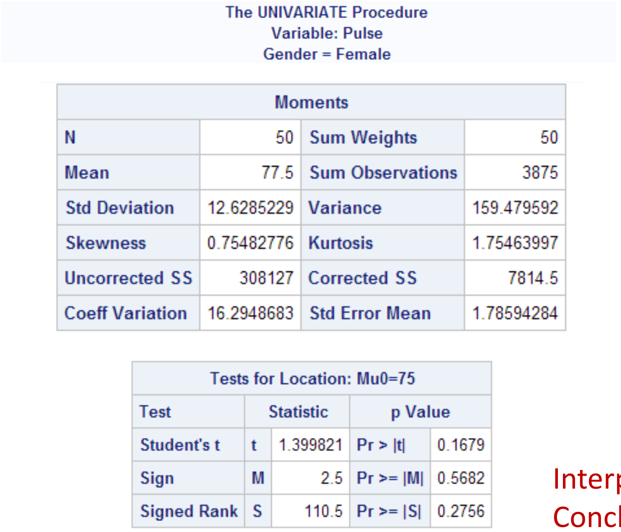
The UNIVARIATE Procedure
Variable: Pulse
Gender = Male

Moments			
N	59	Sum Weights	59
Mean	74.1525424	Sum Observations	4375
Std Deviation	13.758776	Variance	189.303916
Skewness	2.14757261	Kurtosis	11.2131414
Uncorrected SS	335397	Corrected SS	10979.6271
Coeff Variation	18.5546921	Std Error Mean	1.79124006

Tests for Location: Mu0=75				
Test	Statistic		p Value	
Student's t	t	-0.47311	Pr > |t|	0.6379
Sign	M	-0.5	Pr >= |M|	1.0000
Signed Rank	S	-127	Pr >= |S|	0.3166

Interpretation?
Conclusions?

Exercise: Pulse rates by *Gender*



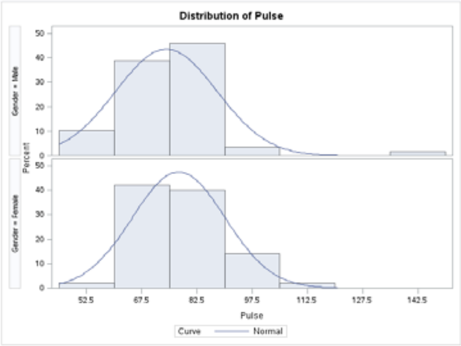
The UNIVARIATE Procedure
Variable: Pulse
Gender = Female

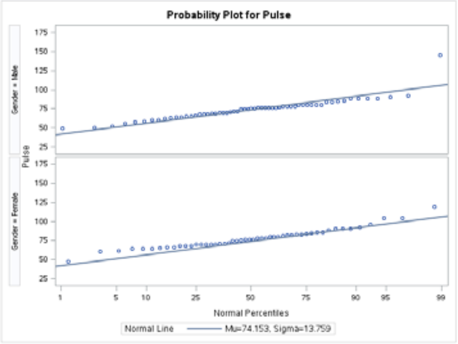
Moments			
N	50	Sum Weights	50
Mean	77.5	Sum Observations	3875
Std Deviation	12.6285229	Variance	159.479592
Skewness	0.75482776	Kurtosis	1.75463997
Uncorrected SS	308127	Corrected SS	7814.5
Coeff Variation	16.2948683	Std Error Mean	1.78594284

Tests for Location: Mu0=75				
Test	Statistic		p Value	
Student's t	t	1.399821	Pr > |t|	0.1679
Sign	M	2.5	Pr >= |M|	0.5682
Signed Rank	S	110.5	Pr >= |S|	0.2756

Interpretation?
Conclusions?

Exercise: Pulse rates by *Gender*





Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.14897825	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.18360906	Pr > W-Sq	0.008
Anderson-Darling	A-Sq	1.35634007	Pr > A-Sq	<0.005

Is a t-test appropriate?

Example: Pulse rates by *Gender*

Using PROC UNIVARIATE:

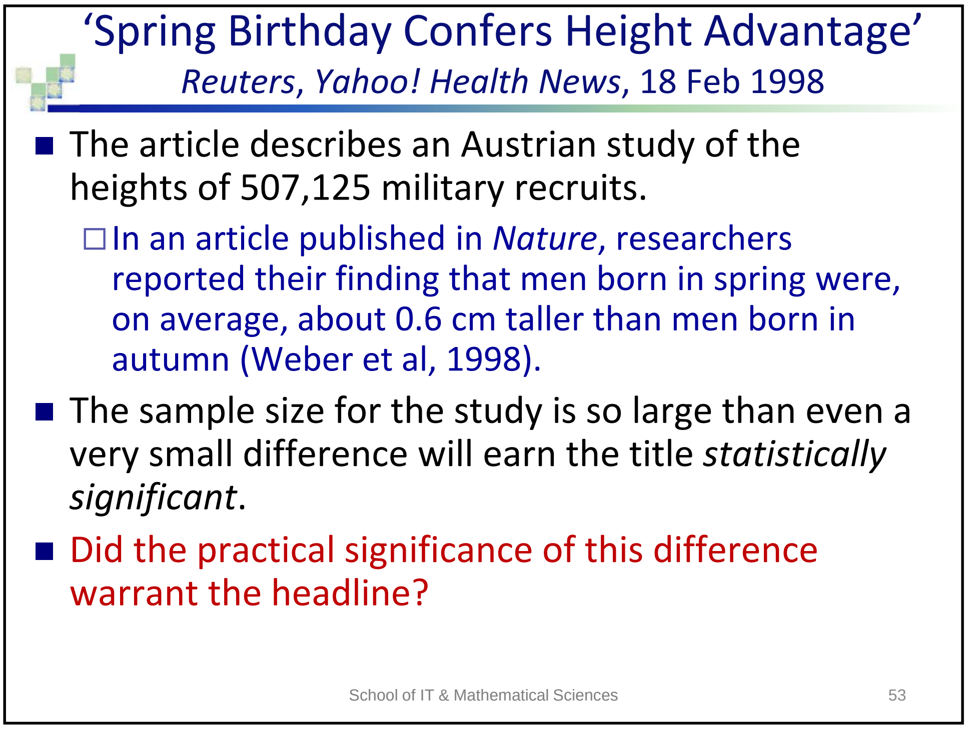
```
proc univariate data=mydata.pulse_rates mu0=75;
    var Pulse;
    format Gender gender.;
    class Gender;
    histogram / normal;
    probplot / normal(mu=est sigma=est);
run;
```

Statistical inference – points to consider

- In many statistical explanations, we use double negatives:
  - □ They are used to communicate that while we are not rejecting a position, we are also not saying it is correct.
- Significance levels should reflect consequence of errors:
  - □ The significance level selected for a test should reflect the consequences associated with Type I and Type II errors.

Statistical inference – points to consider

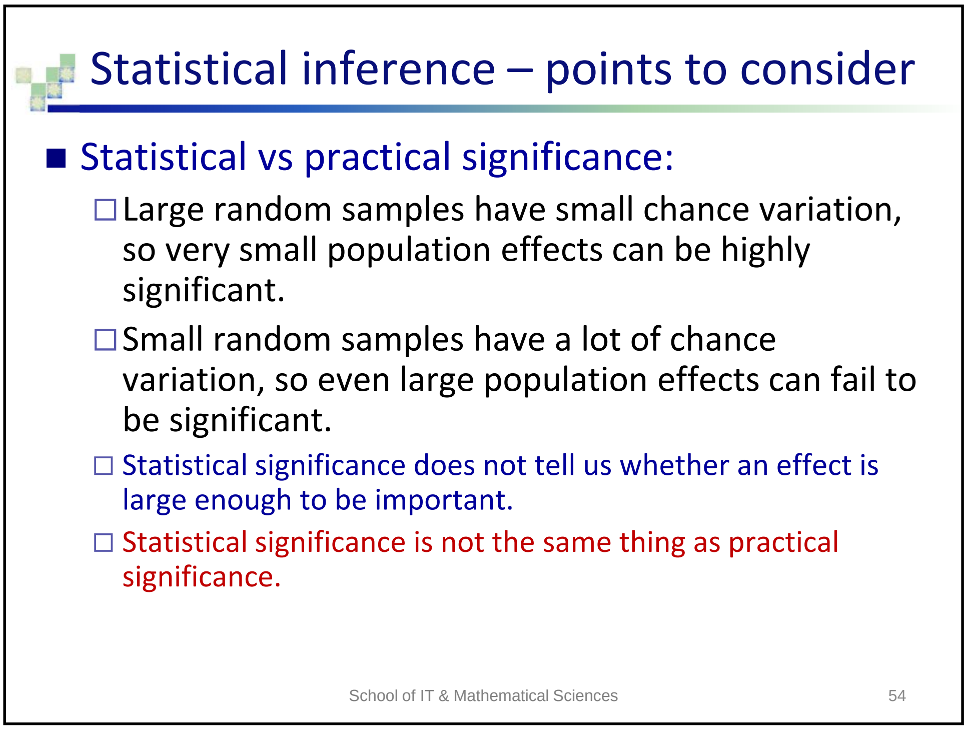
- One-sided vs two-sided hypothesis tests:
  - □ If the researchers are only interested in showing an increase or decrease, but not both, they should use a one-sided test.
  - □ If they would be interested in any difference from the null value, then the test should be two-sided.
  - □ Caution: One-sided hypotheses are allowed only before seeing the data.
    - After observing the data, it is tempting to turn a two-sided test into a one-sided test. Avoid this temptation as it increases the chances of Type I errors.



‘Spring Birthday Confers Height Advantage’

Reuters, *Yahoo! Health News*, 18 Feb 1998

- The article describes an Austrian study of the heights of 507,125 military recruits.
  - □ In an article published in *Nature*, researchers reported their finding that men born in spring were, on average, about 0.6 cm taller than men born in autumn (Weber et al, 1998).
- The sample size for the study is so large than even a very small difference will earn the title *statistically significant*.
- Did the practical significance of this difference warrant the headline?



Statistical inference – points to consider

- Statistical vs practical significance:
  - □ Large random samples have small chance variation, so very small population effects can be highly significant.
  - □ Small random samples have a lot of chance variation, so even large population effects can fail to be significant.
  - □ Statistical significance does not tell us whether an effect is large enough to be important.
  - □ Statistical significance is not the same thing as practical significance.