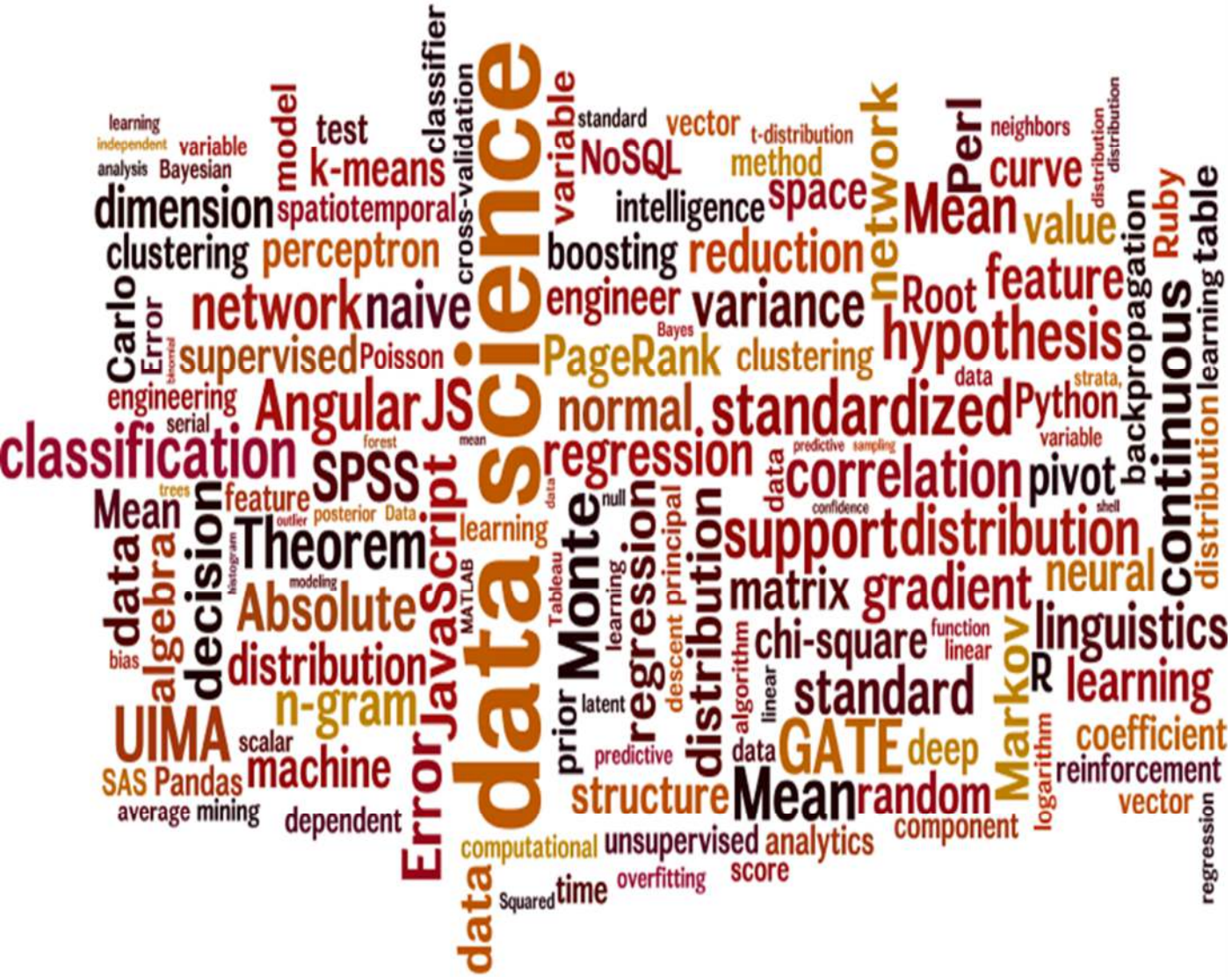


THE UNIVERSITY OF WESTERN AUSTRALIA

STAT2401 Analysis of Experiments

Lecture Week 2

Dr Darfiana Nur

Aims of Lecture Week 2

THE UNIVERSITY OF WESTERN AUSTRALIA

Aim 1 REVIEW 1: Statistical inference (Moore et al 2021, Ch 2)

- **Hypothesis Testing (Test Significance)**
- **Confidence Interval**

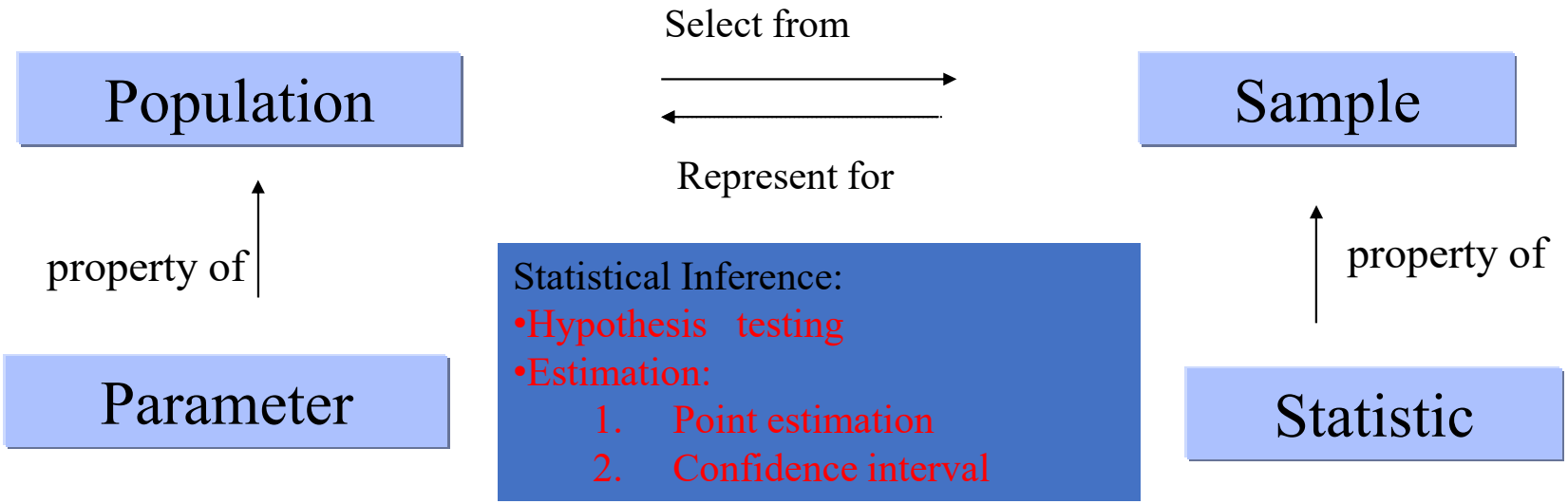
Aim 2 REVIEW 2 The Central Limit Theorem, Normal and t distributions for Sampling distributions

Aim 3 REVIEW 3 Two independent sample t-test

Aim 4 Visualisation

Aim 5 Summary of Hypothesis Testing (independent reading)

Statistical Inference: in general



examples:

μ mean

σ standard deviation

p proportion

examples:

$\bar{x}$

s

$\hat{p}$

Which statistical modelling?

Type of objective	Type of data	Statistical method/model
RELATIONSHIPS	2 numerical (1 numerical explanatory, 1 numerical response)	Linear regression
COMPARISONS	1 categorical (2 sub-categories) – 1 sample	z-test or CI for a proportion
	Numerical – 1 sample from 1 population	One sample t-test or CI
	Numerical – 1 sample paired	Paired t-test or CI
	Numerical – 2 samples from 2 independent populations	Two sample t-test or CI

- 1. Comparing vitamin content of bread immediately after baking vs. 3 days later (the same loaves are used on day one and 3 days later). pair
- 2. Comparing vitamin content of bread immediately after baking vs. 3 days later (tests made on independent loaves). 2 sample
- 3. Average fuel efficiency for 2005 vehicles is 21 miles per gallon. Is average fuel efficiency higher in the new generation “green vehicles”? one sample
- 4. Is blood pressure altered by use of an oral contraceptive? Comparing a group of women not using an oral contraceptive with a group taking it. 2 sam
- 5. Review insurance records for dollar amount paid after fire damage in houses equipped with a fire extinguisher vs. houses without one. Was there a difference in the average dollar amount paid? 2 sample

In Class Exercise 1

Which type of test?
One sample, paired samples, two samples?

AIM 1 REVIEW 1 Statistical Inference

1. Hypothesis Testing
2. Confidence Interval

focusing on 2-sample independent t-test

1. Hypothesis Testing or Test of Significance

6 Steps in carrying out a hypothesis testing: in general

1. State the hypotheses
2. Calculate the test statistic
3. Determine the sampling distribution of the test statistic in (2)
4. Find the p-value based on (3)
5. Make a decision based on (4) p value small - reject null
6. State your conclusion in the context of your specific setting.

What is a Hypothesis?

- A statement about the population parameter in question.

- We generally use two contradictory statements (hypotheses)
  - the null hypothesis, H_0
  - the alternative hypothesis, H_A (or H_1)

STEP 1
Hypotheses: two contradictory statements

- Null hypothesis (H_0): Usually represent the status quo, the no effect is present or the prior belief.

- Alternative hypothesis (H_A): Simply states the case that the parameter differs from its null (H_0) in a specific direction (one sided) or in either direction (two-sided).

- Usually both H_0 and H_A are stated in terms of the population parameters .

EXAMPLE (Moore et al 2021)

Can directed reading activities in the classroom help improve reading ability? A class of 21 third-graders participates in these activities for 8 weeks while a control classroom of 23 third-graders follows the same curriculum without the activities. After 8 weeks, all children take a reading test (scores in table).

Treatment group				Control group			
24	61	59	46	42	33	46	37
43	44	52	43	43	41	10	42
58	67	62	57	55	19	17	55
71	49	54		26	54	60	28
43	53	57		62	20	53	48
49	56	33		37	85	42	

Group	n	$\bar{x}$	s
Treatment	21	51.48	11.01
Control	23	41.52	17.15

Keywords: Can directed reading activities in the classroom help improve reading ability?

Step 1. State the hypotheses

H_0: $\mu_T = \mu_C$ or $\mu_T - \mu_C = 0$

H_A: $\mu_T > \mu_C$

μ_T, μ_C : population reading score means from treatment and control groups respectively

Step 2: Test Statistic

- In hypothesis tests the test statistic summarises the differences between the observed (sample) and expected data (population under the null hypothesis).
- It is a random variable with a distribution that we know.

Step 2 Test statistic

$$t = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_b}}}$$

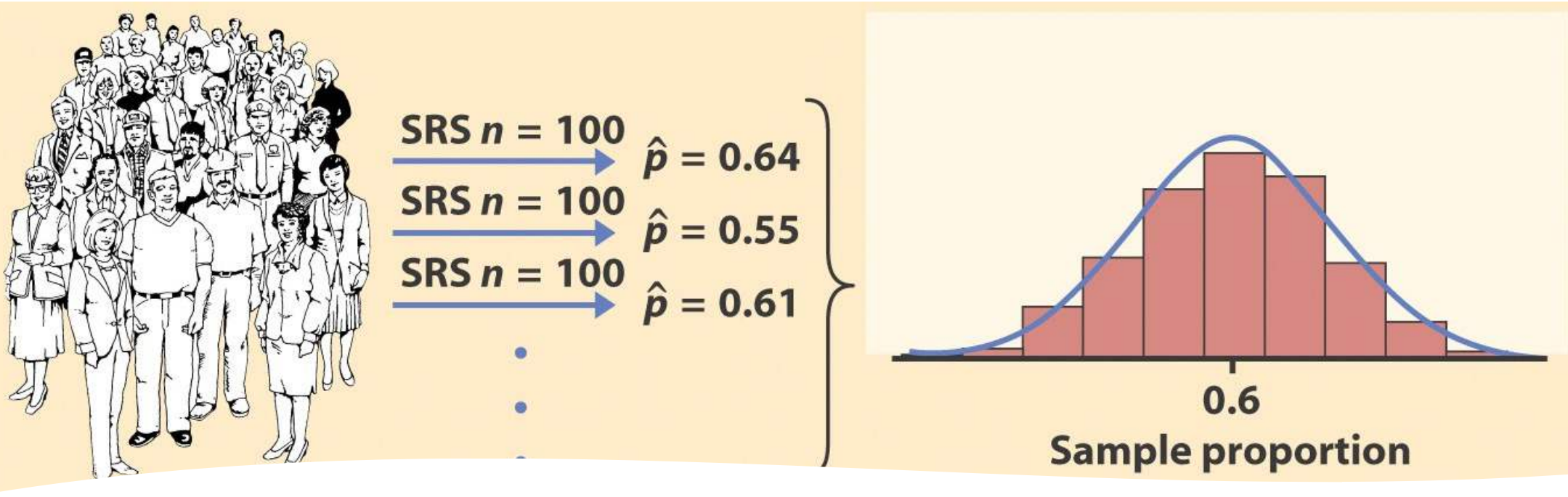
= 2.311 (or see R output under “t stat”)

What is a sampling distribution? (For Step 3)

- The **sampling distribution of a statistic** is the distribution of all possible values taken by **the test statistic** when all possible samples of a fixed size *n* are taken from the population.

- It is a theoretical idea—we do not actually build it.

- The sampling distribution of a statistic is the **probability distribution** of that statistic.



Sampling variability

- Each time we take a random sample from a population, we are likely to get a different set of individuals and calculate a different statistic. This is called **sampling variability.**

- If we take a lot of random samples of the same size from a given population, the variation from sample to sample—the **sampling distribution**—will follow a predictable pattern.

Keywords: Can directed reading activities in the classroom help improve reading ability?

Step 1. State the hypotheses

H_0: $\mu_T = \mu_C$ or $\mu_T - \mu_C = 0$

H_A: $\mu_T > \mu_C$

Step 2 Test statistic

$$t = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_b}}}$$

= 2.311 (or see R output under "t stat")

μ_T, μ_C : population reading score means from treatment and control groups respectively

STEP 3 Sampling Distribution

This test statistic approximately follows a **Student's *t* distribution with df ≈ 38**

t ~ T with df

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Step 4: p-value

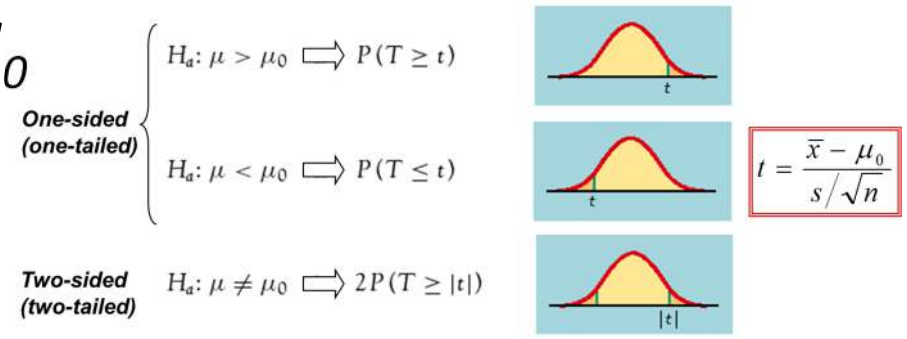
The test statistic helps us find the probability of getting an outcome:

"as extreme as, or more extreme than, the actual observed outcome".

"extreme" means "far" from what we would expect if the null was true.

What is a p-value?

- The p-value can be interpreted as the strength of the evidence provided by the observed data against H_0 .
- The ***p-value*** is the probability, if H_0 is true, of randomly drawing a sample like the one obtained or more extreme, in the direction of H_A.
- p-value must always be between 0 and 1.
- small p-value: strong evidence against H_0
  large p-value: weak evidence against H_0



Step 5: Decision

- If the p-value is small, reject H_o
- If the p-value is large, do not reject H_o

Decision based on a Significance Level

- We can also compare the obtained *p*-value with a fixed value that the researcher regards as decisive.
- Meaning that you need to state in advance how much evidence against the null you will require to reject the null.
- This fixed value is called the significance level and usually known as α

Decision based on a significance level

- The decision requires us to compare the p-value we found from the test statistic with the significance level α .

If p-value > significance level α
do not reject H_o.

If p-value $\leq$ significance level α

reject H_o.

If the p-value is as small or smaller than α , we say that the data is statistically significant at level α

Steps 4 - 6

Step 4. p-value [based on Ha: $\mu_T > \mu_C$]

p-value = $P(t_{38} > 2.311) = 0.013$ *(using R, a half of the p-value for two sided)*

Steps 5 and 6. Decision and Conclusion:

This small p-value=0.013 =1.3% < α=5% strongly suggests that the data provides very strong evidence against H_0 (therefore **we do reject H_o**);

We conclude that there is statistically significant difference in the mean reading score of treatment and control groups.

2. Confidence interval (CI) for a population mean

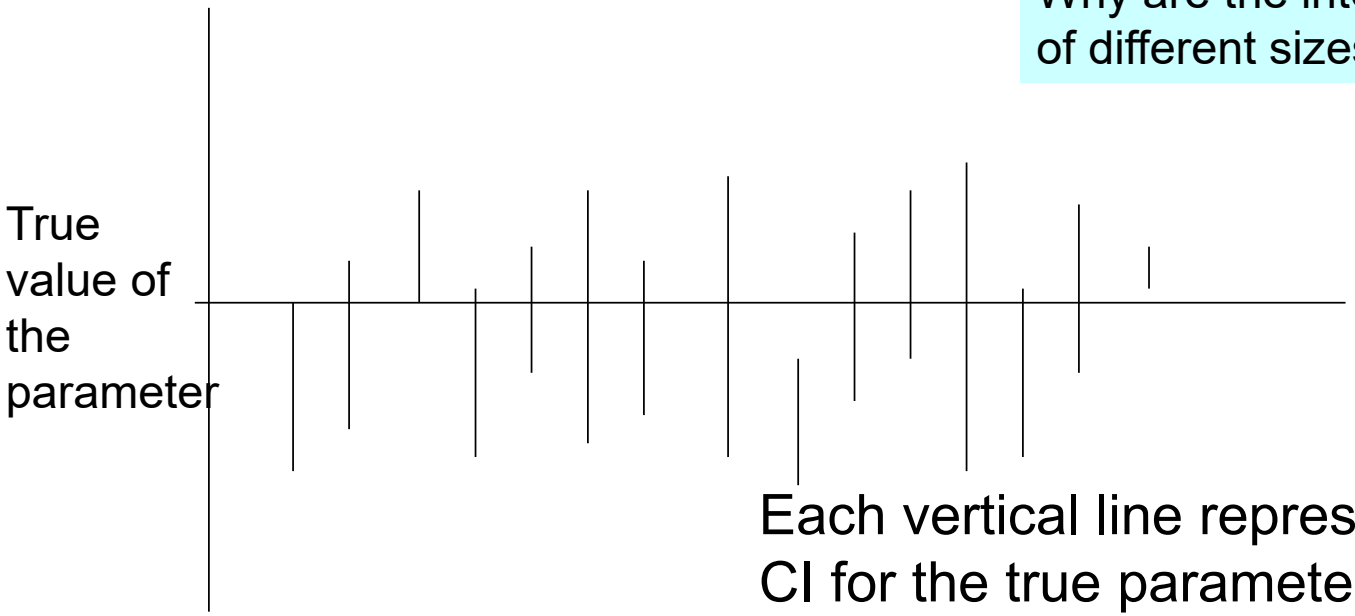
A confidence interval is used when we wish to

estimate a population parameter

rather than test a hypothesis about the population.

If repeated samples were taken and the CIs were plotted (vertically) they may look like this . . .

Why are the intervals of different sizes?



Each vertical line represents a CI for the true parameter - one CI per sample drawn.

Here you would expect CI% intervals to contain the true mean.

Confidence interval (CI)

sample mean from the sample

CI = point estimate ± margin of error (m)

CI for $(\mu_A - \mu_B)$: $(\bar{x}_A - \bar{x}_B) \pm t^* \sqrt{\frac{{s_A}^2}{n_A} + \frac{{s_B}^2}{n_B}}$

m = multiplier × standard error

Multiplier t* is defined based on the sampling distribution

AIM 2

REVIEW 2

Central Limit Theorem and Distributions for Sampling Distributions

The Central Limit Theorem

- Draw a Simple Random Sample of size *n*, from any population with a mean and finite standard deviation.

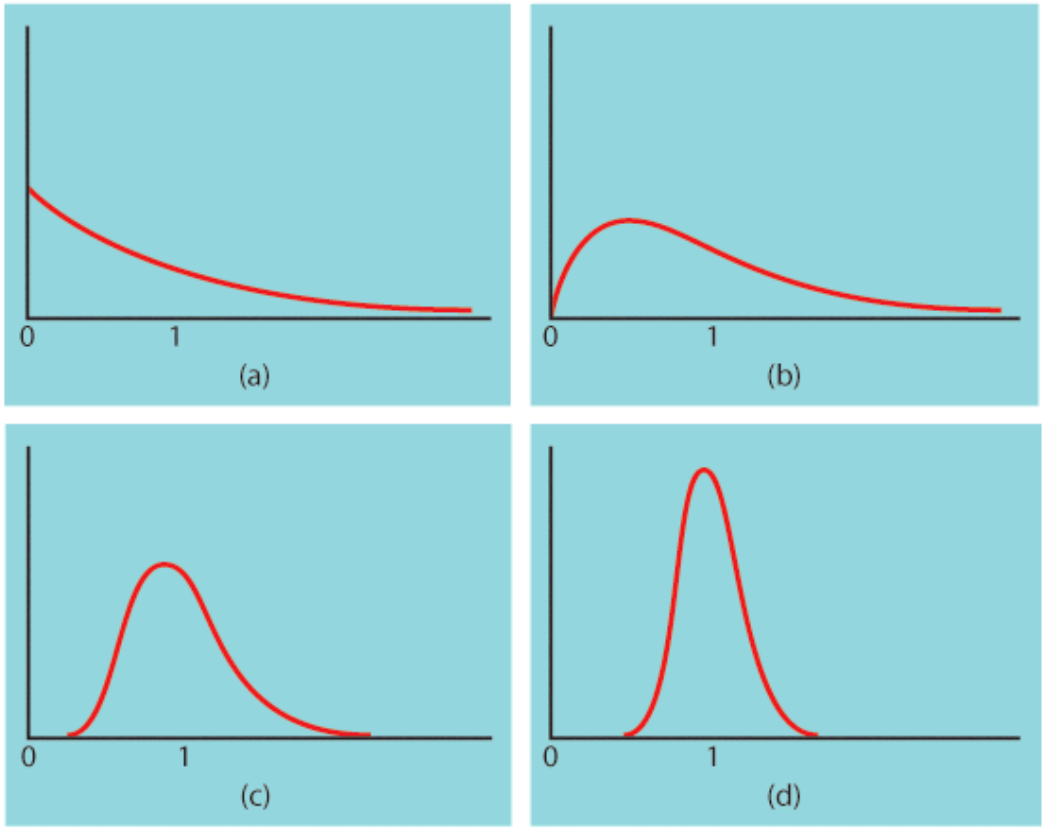
- When *n* is large, the sampling distribution of the sample mean is:

$$\bar{X} \sim N\left(\mu_{\bar{X}} = \mu_X, \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}\right)$$

The central limit theorem

Central Limit Theorem (CLT): When randomly sampling from any population with population mean μ and population standard deviation σ, when *n* is large enough, the sampling distribution of sample mean $\bar{x}$ is approximately Normal: $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$.

Population with strongly skewed distribution



Sampling distribution of $\bar{x}$ for *n* = 2 observations

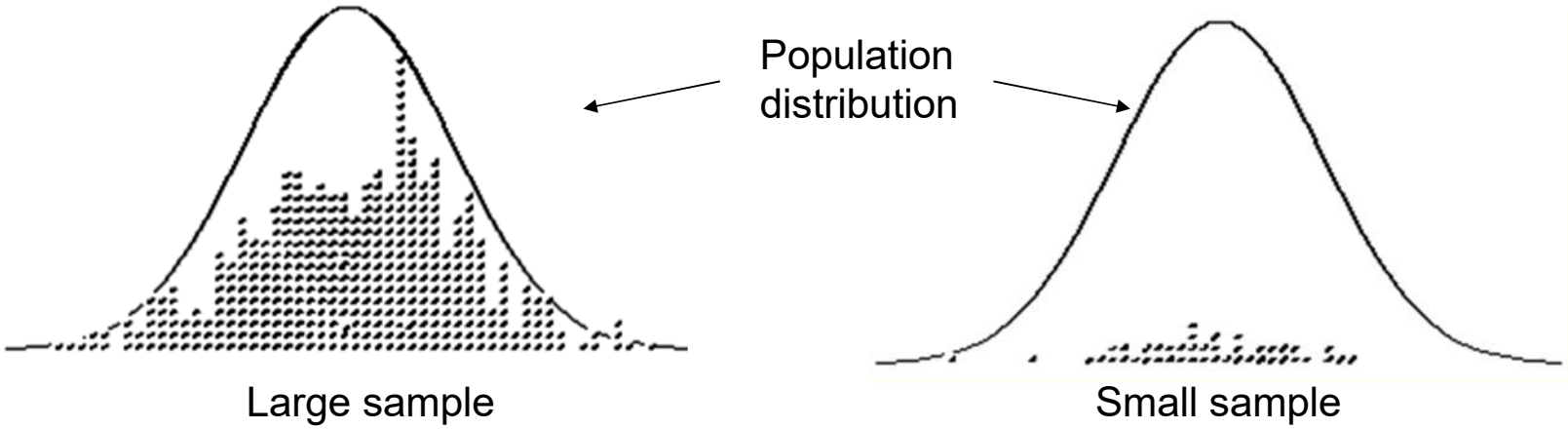
Sampling distribution of $\bar{x}$ for *n* = 10 observations

Sampling distribution of $\bar{x}$ for *n* = 25 observations

When σ is unknown

The sample standard deviation *s* provides an estimate of the population standard deviation σ.

- When the sample size is large, the sample is likely to contain elements representative of the whole population. Then *s* is a good estimate of σ.
- But when the sample size is small, the sample contains only a few individuals. Then *s* is a mediocre estimate of σ.



Standard deviation *s* – standard error *s/√n*

For a sample of size *n*,

the sample standard deviation *s* is:

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

n − 1 is the "degrees of freedom."

The value ***s/√n*** is called the standard error of the mean **(SEM).**

Scientists often present sample results as mean ± SEM.

A study examined the effect of a new medication on the seated systolic blood pressure. The results, presented as mean ± SEM for 25 patients, are 113.5 ± 8.9.

What is the standard deviation *s* of the sample data?

SEM = $s/\sqrt{n}$ <=> $s = \text{SEM}*\sqrt{n}$

$s = 8.9*\sqrt{25} = 44.5$

The *t* distributions

Suppose that an SRS of size *n* is drawn from an $N(\mu, \sigma)$ population.

- When σ is known, the sampling distribution is $N(\mu, \sigma/\sqrt{n})$.

- When σ is estimated from the sample standard deviation *s*, the sampling distribution follows a **t distribution $t(\mu, s/\sqrt{n})$ with degrees of freedom $n - 1$.**

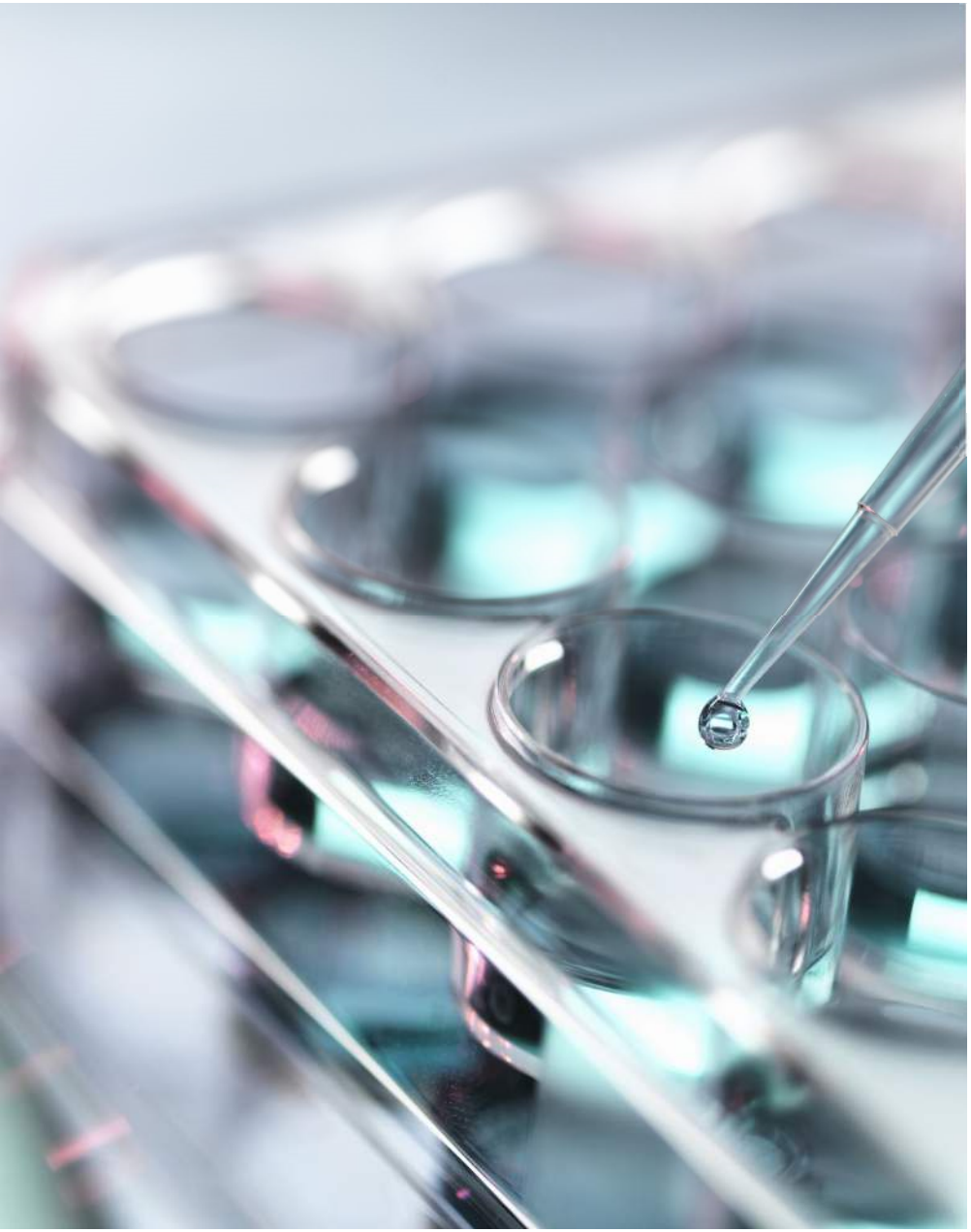
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

is the **one-sample *t* statistic.**

AIM 3

REVIEW 3

Two sample *t* test – independent samples



Two sample *t* test – independent samples

- Comparing random samples that were randomly selected from 2 populations is a two-sample problem.
- The two samples may differ in sample size
- If both sample means are Normally distributed, then we can do a comparison of the means.

Two-sample *z* statistic , σ_1 , σ_2 are known

- We have **two independent SRSs** (simple random samples) possibly coming from two distinct populations with (μ_1, σ_1) and (μ_2, σ_2). We use $\bar{x}_1$ and $\bar{x}_2$ to estimate the unknown μ_1 and μ_2.

- When both populations are normal, the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$ is also normal, with standard deviation : $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Then the **two-sample *z* statistic**

has the standard normal *N(0, 1)*

sampling distribution.

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Two independent samples *t* distribution

- We have **two independent SRSs** (simple random samples) possibly coming from two distinct populations with (μ_1, σ_1) and (μ_2, σ_2) unknown.
- We use $(\bar{x}_1, s_1)$ and $(\bar{x}_2, s_2)$ to estimate (μ_1, σ_1) and (μ_2, σ_2), respectively.
- To compare the means, both populations should be normally distributed. However, in practice, it is enough that the two distributions have similar shapes and that the sample data contain no strong outliers.

Unequal Variances

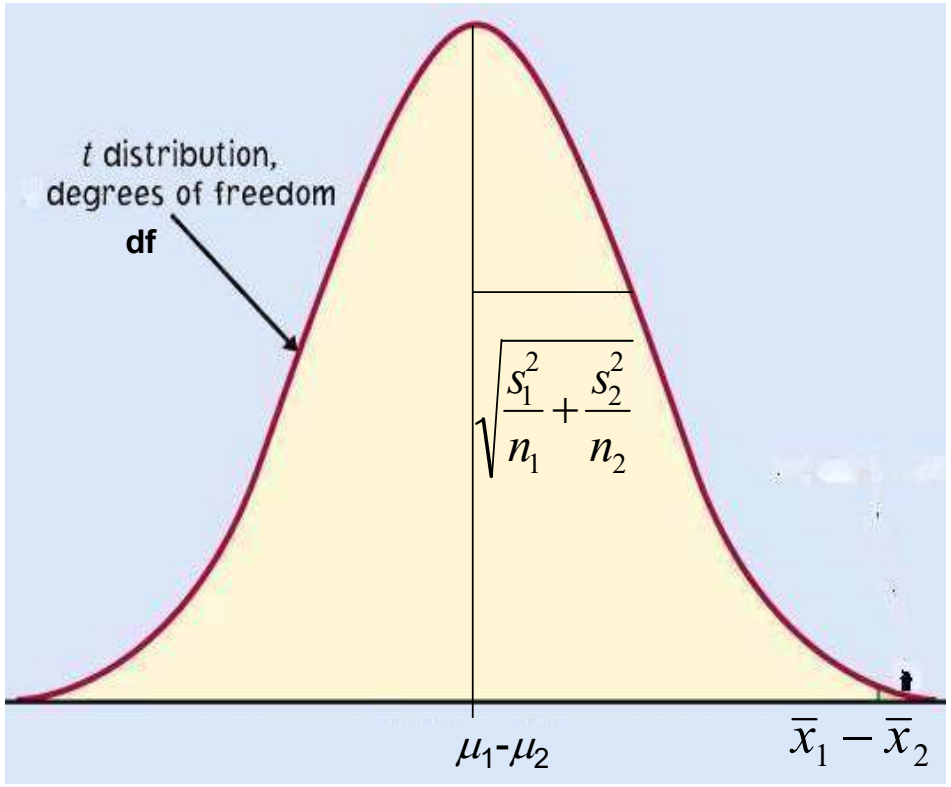
The two-sample *t* statistic follows approximately the t distribution with a standard error SE (spread) reflecting variation from both samples:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Conservatively, the degrees of freedom is equal to the df=smallest of ($n_1 - 1$, $n_2 - 1$).

[Approximation]



Two-sample *t* significance test

The null hypothesis is that both population means μ_1 and μ_2 are equal, thus their difference is equal to zero.

$$H_0: \mu_1 = \mu_2 \iff \mu_1 - \mu_2 = 0$$

with either a one-sided or a two-sided alternative hypothesis.

We find how many standard errors (SE) away from $(\mu_1 - \mu_2)$ is $(\bar{x}_1 - \bar{x}_2)$ by standardizing with *t*:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE}$$

Because in a two-sample test H_0 poses $(\mu_1 - \mu_2) = 0$, we simply use

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

with df = rounded of

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Further details of the two sample *t* procedures – unequal variances

- The **true value of the degrees of freedom** for a two-sample *t*-distribution **is quite lengthy to calculate.**

- That's why some references use an approximate value,

df = smallest($n_1 - 1$, $n_2 - 1$),

which errs on the conservative side (often smaller than the exact).

- Computer software such as R, Excel, JMP gives the exact degrees of freedom—or the rounded value— for your sample data.

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

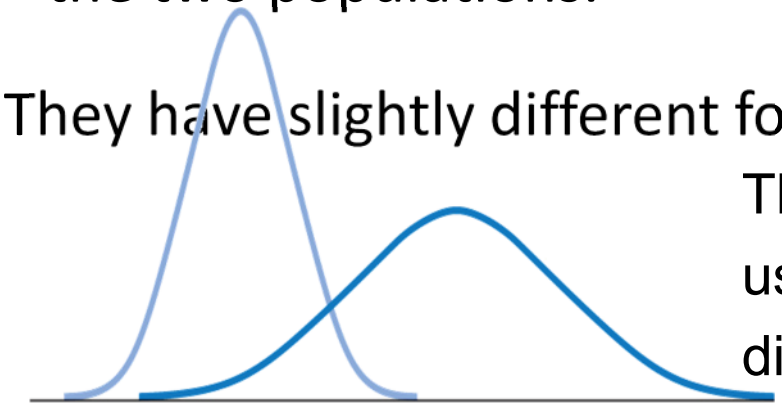
Pooled two-sample procedures

Equal variances

There are two versions of the two-sample *t*-test:

- one **assuming equal variance ("pooled 2-sample test")** and
- one **not assuming equal variance ("unequal" variance,** as we have studied) for the two populations.

They have slightly different formulas and degrees of freedom.



Two normally distributed populations with unequal variances

The pooled (equal variance) two-sample *t*-test was often used before computers because it has exactly the *t* distribution for degrees of freedom $n_1 + n_2 - 2$.

However, the assumption of equal variance is hard to check, and thus the unequal variance test is safer.

When both population have the *same* standard deviation, the **pooled estimator of σ^2** is:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

The sampling distribution for $(x_1 - x_2)$ has exactly the *t* distribution with ***($n_1 + n_2 - 2$) degrees of freedom***.

A level *C* confidence interval for $\mu_1 - \mu_2$ is $(\bar{x}_1 - \bar{x}_1) \pm t * \sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}$

(with area C between $-t^*$ and t^*)

$$t = \frac{(\bar{x}_1 - \bar{x}_1)}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

To test the hypothesis H_0: $\mu_1 = \mu_2$ against a one-sided or a two-sided alternative, compute the pooled two-sample *t* statistic for the $t(n_1 + n_2 - 2)$ distribution.

Two-sample *t* confidence interval

Because we have two independent samples we use the difference between both sample averages ($\bar{x}_1 - \bar{x}_2$) to estimate ($\mu_1 - \mu_2$).

Practical use of *t*: *t**

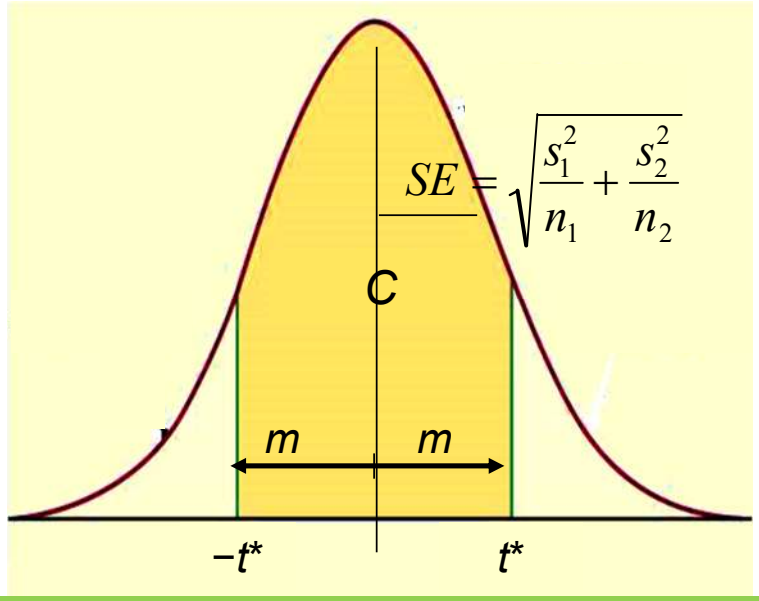
- *C* is the area between $-t^*$ and t^*.
- We find t^* in the line of Table D or R for df = round

$$\left(\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} \right)$$

and the column for confidence level C.

- The margin of error *m* is:

$$m = t * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = t * SE$$



CI for $(\mu_A - \mu_B)$: $(\bar{x}_A - \bar{x}_B) \pm t^* \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$

EXAMPLE (Moore et al 2021)

Can directed reading activities in the classroom help improve reading ability? A class of 21 third-graders participates in these activities for 8 weeks while a control classroom of 23 third-graders follows the same curriculum without the activities. After 8 weeks, all children take a reading test (scores in table).

Treatment group				Control group			
24	61	59	46	42	33	46	37
43	44	52	43	43	41	10	42
58	67	62	57	55	19	17	55
71	49	54		26	54	60	28
43	53	57		62	20	53	48
49	56	33		37	85	42	

Group	n	$\bar{x}$	s
Treatment	21	51.48	11.01
Control	23	41.52	17.15

$$CI: (\bar{x}_1 - \bar{x}_2) \pm m; m = t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

95% confidence interval for the reading ability study using the more precise degrees of freedom:

$$\mathrm{df} = \frac{\left(\frac{11.01^2}{21} + \frac{17.15^2}{23}\right)^2}{\frac{1}{20}\left(\frac{11.01^2}{21}\right)^2 + \frac{1}{22}\left(\frac{17.15^2}{23}\right)^2}$$

$$= \frac{344.486}{9.099} = 37.86$$

30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423
	50%	60%	70%	80%	90%	95%	96%	98%
	Confidence level C							

$$m = t * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$m = 2.024 * 4.31 \approx 8.72$$

t-Test: Two-Sample Assuming Unequal Variances

	Treatment group	*Control group*
Mean	51.476	41.522
Variance	121.162	294.079
Observations	21	23
Hypothesized Mean Difference	-	
df	38	
t Stat	2.311	
P(T<=t) one-tail	0.013	
t Critical one-tail	1.686	
P(T<=t) two-tail	0.026	
t Critical two-tail	2.024	

Independent Samples Test

		Levene's Test for Equality of Variances		*t*-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Reading Score	Equal variances assumed	2.362	.132	2.267	42	.029	9.95445	4.39189	1.09125	18.81765
	Equal variances not assumed			2.311	37.855	.026	9.95445	4.30763	1.23302	18.67588

SPSS

Summary

Keywords: Can directed reading activities in the classroom help improve reading ability?

μ_T, μ_C : population reading score means from treatment and control groups respectively

Step 1. State the hypotheses

H_0: $\mu_T = \mu_C$ or $\mu_T - \mu_C = 0$

H_A: $\mu_T > \mu_C$

Step 2 Test statistic

$$t = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_b}}}$$

= 2.311 (or see R output under "t stat")

STEP 3 Sampling Distribution

This test statistic approximately follows a **Student's *t* distribution with df ≈ 38**

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Steps 4 - 6

Step 4. p-value

p-value $= P(t_{38} > 2.311) = 0.013$ *(using R, a half of the p-value for two sided)*

Steps 5 and 6. Decision and Conclusion:

This small p-value=0.013 =1.3% < α=5% strongly suggests that the data provides very strong evidence against H_0 (therefore **we do reject H_o**);

We conclude that there is statistically significant difference in the mean reading score of treatment and control groups.

Aim 3 Multivariate Data – Visualisation

- With the advances of technology, the question is no longer "shall we plot?" but rather "what shall we plot?"
- Scatterplot, boxplot and histogram
  - Use scatterplot for bivariate continuous variables
  - Use boxplot/histogram for the marginal distribution or the distribution of one numerical data

In-Class Exercise 3.

Interpret the boxplot.

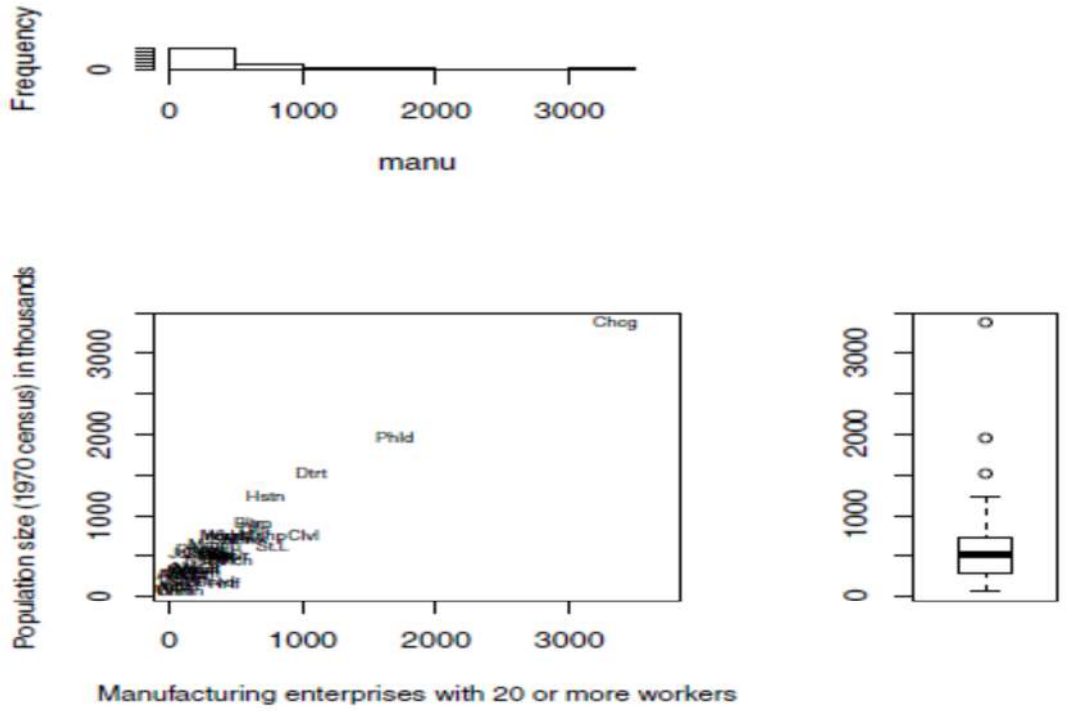


Fig. 2.3. Scatterplot of manu and popul that shows the marginal distributions by histogram and boxplot.

Multivariate Data – Visualisation EH(2011) Ch 2

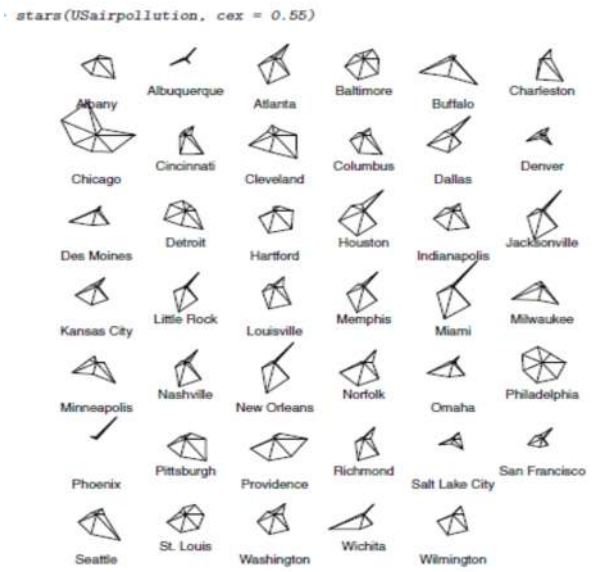


Fig. 2.9. Star plot of the air pollution data.

- The bubble plot: for 3 variables three variables are displayed; two are used to form the scatterplot itself, and then the values of the third variable are represented by circles with radii proportional to these values and centred on the appropriate point in the scatterplot.
- Star and Chernoff face plots: representations of each multivariate observation to group observations on the basis of similarities.

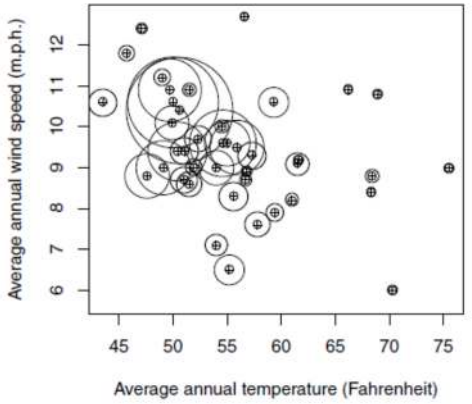
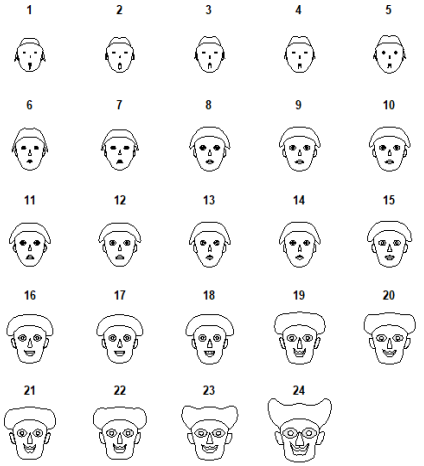


Fig. 2.7. Bubble plot of temp, wind, and SO2.

The cities with moderate annual temperatures and moderate annual wind speeds tend to suffer the greatest air pollution.



Variable	Facial characteristic
X_1: Fixed-charge coverage	↔ Half-height of face
X_2: Rate of return on capital	↔ Face width
X_3: Cost per kW capacity in place	↔ Position of center of mouth
X_4: Annual load factor	↔ Slant of eyes
X_5: Peak kWh demand growth from 1974	↔ Eccentricity $\left(\frac{\text{height}}{\text{width}}\right)$ of eyes
X_6: Sales (kWh use per year)	↔ Half-length of eye
X_7: Percent nuclear	↔ Curvature of mouth
X_8: Total fuel costs (cents per kWh)	↔ Length of nose

Multivariate Data – Visualisation EH(2011) Ch 2

- A bivariate boxplot: which is a two-dimensional analogue of the boxplot for univariate to identify outliers in a scatterplo
- The chi-plot: to check the independence between two variables

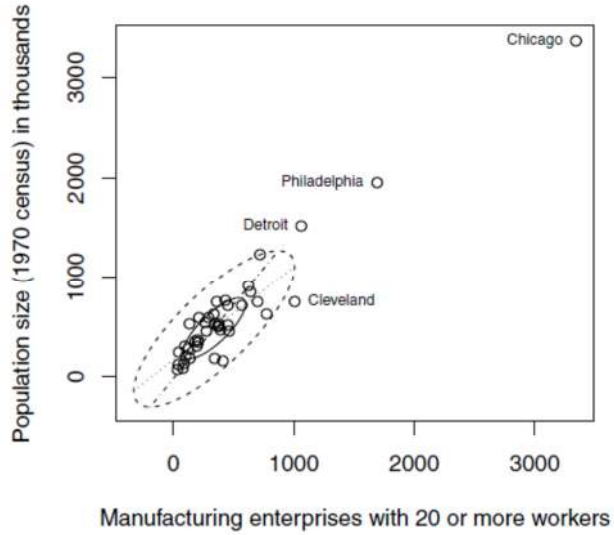


Fig. 2.4. Scatterplot of manu and popul showing the bivariate boxplot of the data.

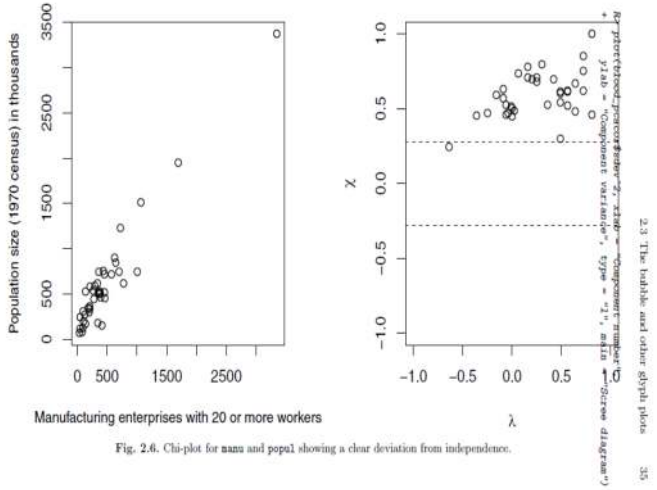


Fig. 2.6. Chi-plot for manu and popul showing a clear deviation from independence.

Chicago, Philadelphia, Detroit, and Cleveland are outliers

but not Houston, because it is on the fence rather than outside the fence.

Departure from independence is indicated in the chi plot by a lack of points in the horizontal band indicated on the plot. Here there is a very clear departure since there are very few of the observations in this region.

Multivariate Data - Visualisation

Fig 2.11 The scatterplot matrix shows the presence of possible outliers in many panels and the relationship between the precip, predays, and SO2 might be non-linear.

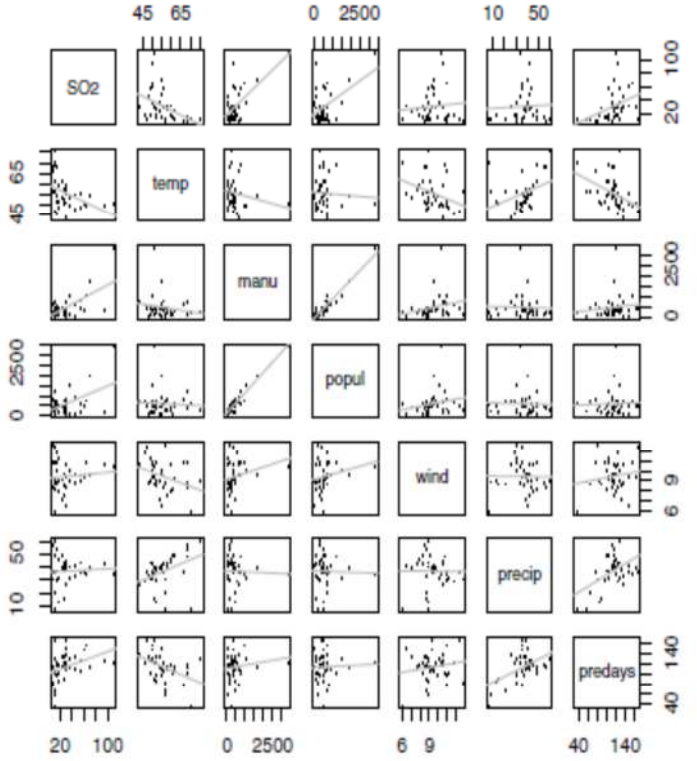


Fig. 2.11. Scatterplot matrix of the air pollution data showing the linear fit of each pair of variables.

Scatterplot matrix: for multivariate continuous variables

- to identify outliers
- correlation and covariance
- to identify regions of high or low densities of observations that may indicate the presence of distinct groups of observations ie clusters.

Fig 2.17 The waist/hips panel gives some evidence that there might be two groups, being men and women. The Waist histogram on the diagonal panel is also bimodal, underlining the two-group nature of the data.

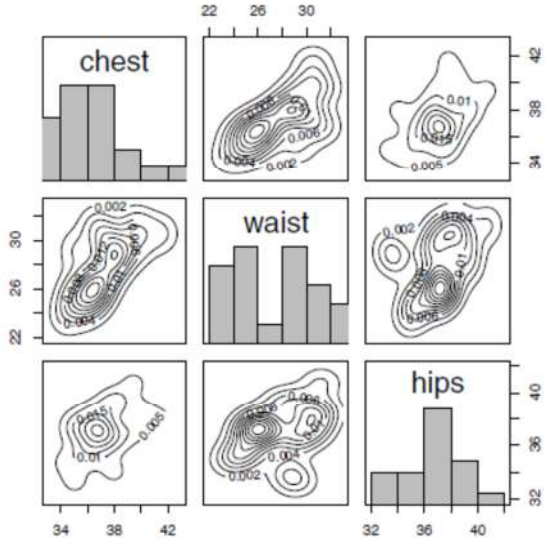
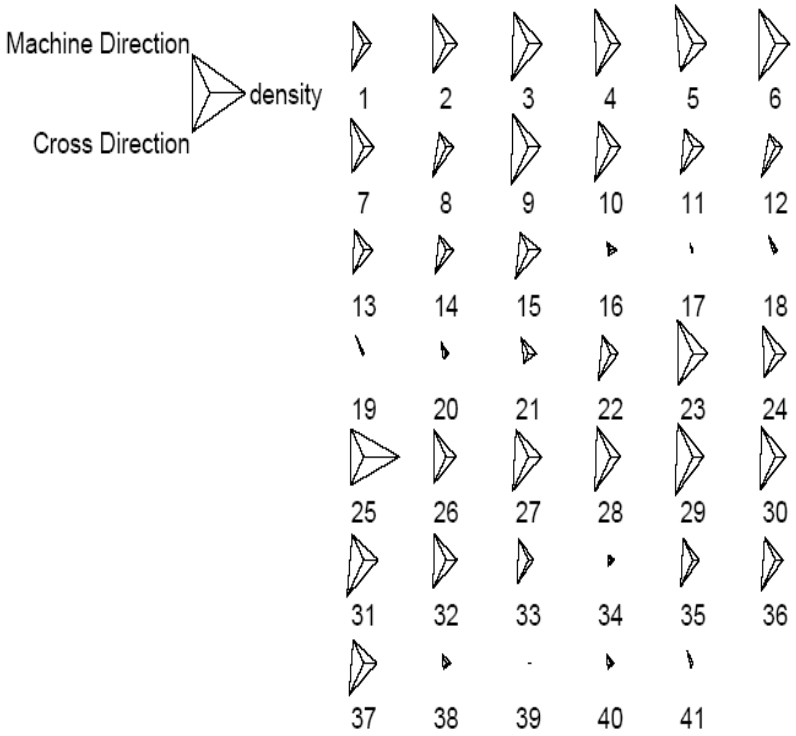
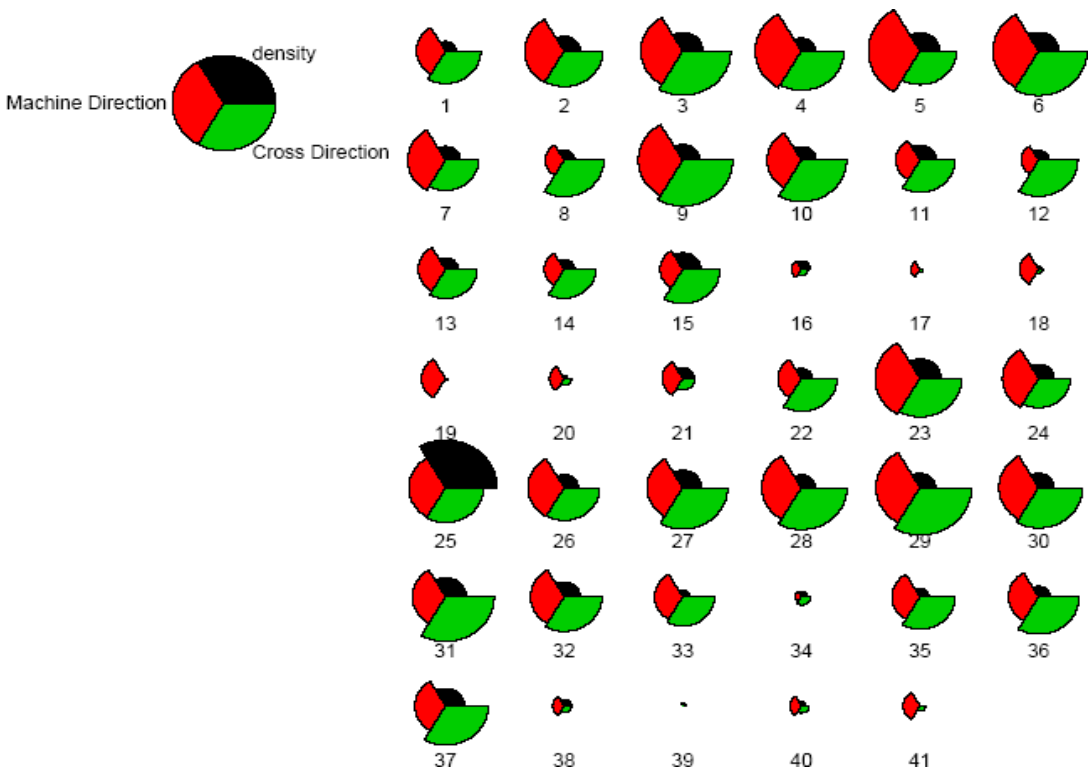


Fig. 2.17. Scatterplot matrix of body measurements data showing the estimated bivariate densities on each panel.

Star plot and segment plots using R



R Command
stars(datat1.2, key.loc=c(-1,15))



R Command
stars(datat1.2,draw.segment=T, key.loc=c(-2,15))

Three or more dimensional plots

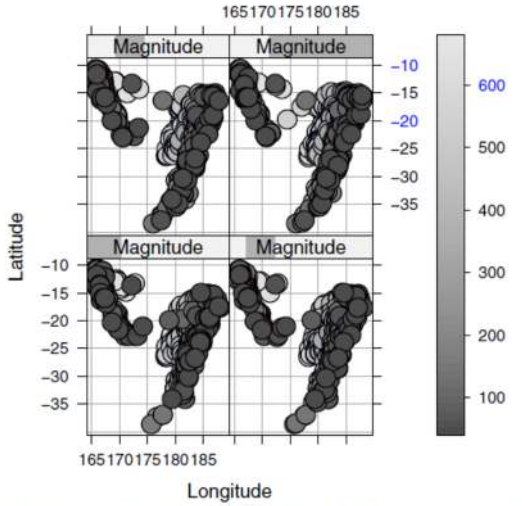


Fig. 2.23. Scatterplots of latitude and longitude conditioned on magnitude, with depth coded by shading.

```
R> with(USairpollution,
+      scatterplot3d(temp, wind, SO2, type = "h",
+                    angle = 55))
```

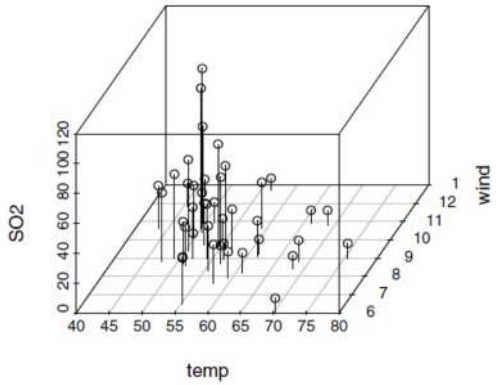


Fig. 2.19. A three-dimensional scatterplot for the air pollution data.

```
R> pollution <- with(USairpollution, equal.count(SO2,4))
R> plot(cloud(precip ~ temp * wind | pollution, panel.aspect = 0.9,
+       data = USairpollution))
```

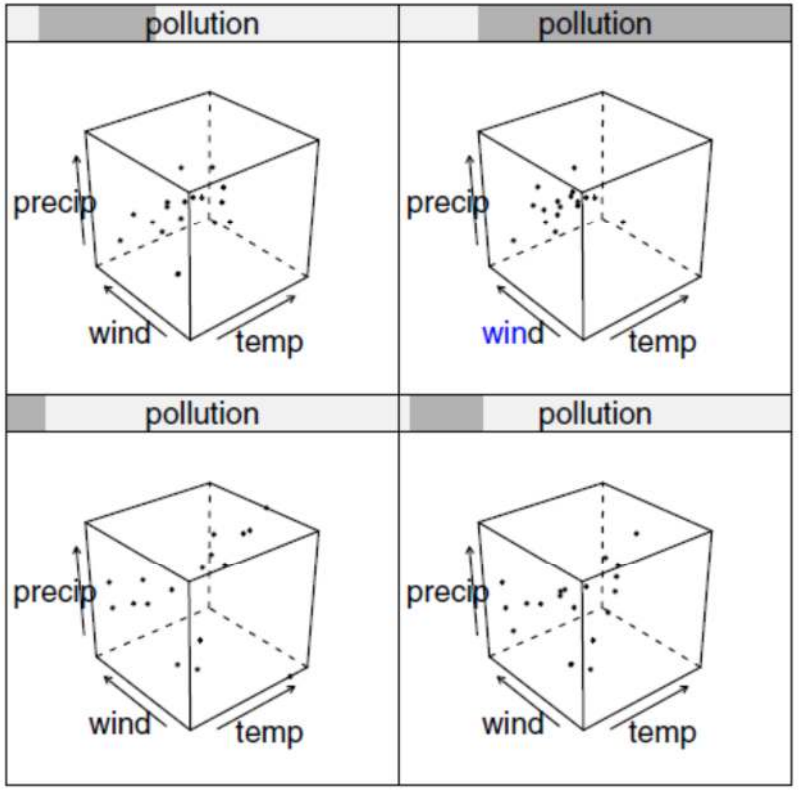
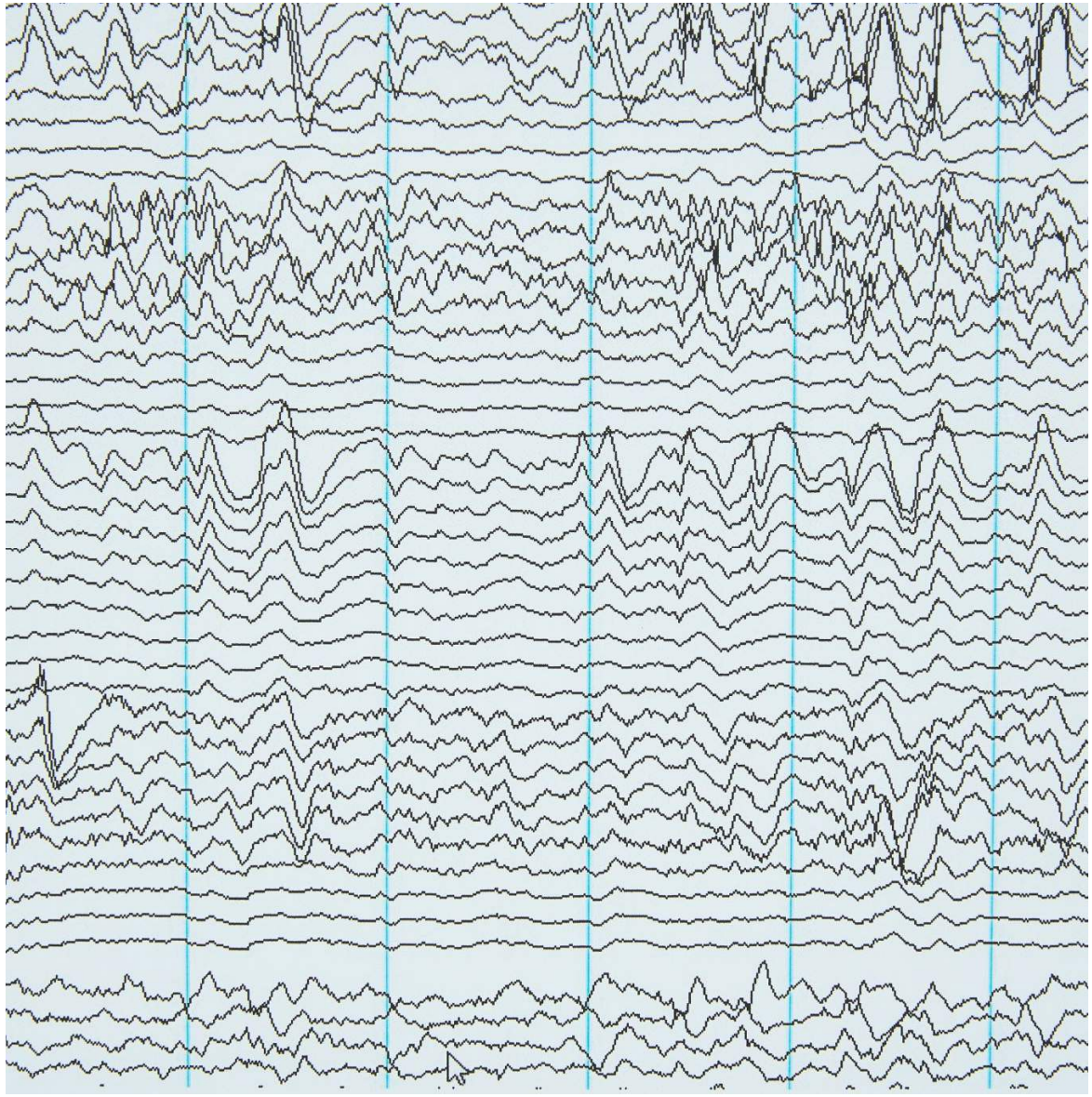


Fig. 2.21. Three-dimensional plots of `temp`, `wind`, and `precip` conditioned on levels of `SO2`.

Graphical Identification of Outliers

- Visually inspect for the unusual observation
  - Use scatterplot of two variables for every two variables.
  - Use scatterplot matrix.
  - Use boxplot for individual variables.
- Look at the correlation coefficient with and without the unusual observation.
- Use three dimensional scatter plots of observed data and also for standardized data.
- Look for visual explanations of the variation from the above plots.
- Use Star and Chernoff face representations of each multivariate observation to group observations on the basis of similarities.



SUMMARY
(Independent reading)

6 Steps in carrying out a hypothesis test

1. State the hypotheses (Ho and Ha)
2. Calculate the test statistic
3. Sampling distribution of the test statistic
4. Find the p-value based on (3), look at Ha (one sided or two sided)
5. Make a decision based on the p-value

- p-value $\leq \alpha$, reject Ho;
- p-value $> \alpha$, do not reject Ho

6. State your conclusion in the context of your specific setting.

Large-sample significance test for a Population Proportion:
Two-sided test

STEP 1 H_0: $p=p_0$ H_A: $p \neq p_0$

STEP 2 The test statistic is: $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \sim N(0,1)$

STEP 3 Normal approximation to Binomial: Z ~ N(0,1)

STEP 4 p-value

p-value = P(|Z |> z) = P(Z < -z) + P(Z > z)

STEP 5 Decision

STEP 6 Conclusion

How about ONE-SIDED *t*-test

[H_A: $p>p_0$ OR H_A: $p<p_0$] - p-value ?

Assumptions/Conditions for 1 sample z-test for one proportion?

Two-sided *t*-test: 1 sample t-test

Step 1: Hypotheses

H_0: $\mu = 0$

H_A: $\mu \neq 0$

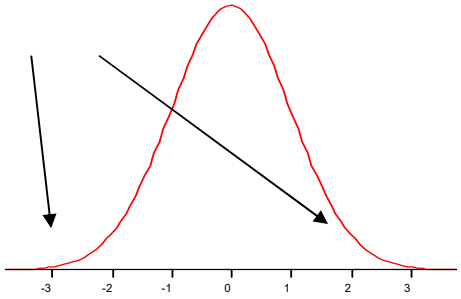
Steps 2 and 3: Test statistic

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

Step 4: Determine the p-value

p-value = probability that a random variable having the $t_{(\mathrm{df}=n-1)}$ distribution exceeds t (in absolute terms)

ie $2P(t_{n-1} > |t|)$



Step 5 Decision

Step 6 Conclusion

How about ONE-SIDED *t*-test

[H_A: $\mu > 0$ OR H_A: $\mu < 0$] - p-value ?

Assumptions for 1 sample *t*-test?

Hypotheses for paired sample *t*-test

1) $H_0: \mu_d = 0$ $H_A: \mu_d \neq 0$

2) Test statistic:

$$t = \frac{\bar{x}_{\text{diff}} - \mu_{\text{diff}}}{\frac{s_{\text{diff}}}{\sqrt{n}}}$$

3) The sampling distribution of the test statistic: $t_{(df=n-1)}$

4) The p-value of the *t*-test is the probability that a random variable having the $t_{(df=n-1)}$ distribution exceeds *t* (in absolute terms)

5) Decision

6) Conclusion

How about ONE-SIDED paired *t*-test

[$H_A: \mu_d > 0$ OR $H_A: \mu_d < 0$] - p-value ?

How about the **assumptions** for paired *t*-test?

Hypotheses for two independent sample *t*-test

1) $H_0: \mu_A = \mu_B$ $H_A: \mu_A \neq \mu_B$

2) Test statistic:

$$t = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

$$df = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)^2}{\frac{1}{n_A - 1}\left(\frac{s_A^2}{n_A}\right)^2 + \frac{1}{n_B - 1}\left(\frac{s_B^2}{n_B}\right)^2}$$

3) The sampling distribution of the test statistic: $t_{(df)}$

4) The p-value of the *t*-test is the probability that a random variable having the $t_{(df)}$ distribution exceeds *t* (in absolute terms)

5) Decision

6) Conclusion

How about ONE-SIDED test for 2 independent sample *t*-test:

[$H_A: \mu_A - \mu_B > 0$ OR $H_A: \mu_A - \mu_B < 0$] - p-value?

How about **the assumptions** for 2 independent sample *t*-test?

Confidence interval (CI)

CI = point estimate ± margin of error (ME)

ME = multiplier × standard error

INTERPRETATIONS

Hypothesis Testing and CI are consistent for the same significance level

CI for 1−sample z−test (proportion) : $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

CI for 1−sample t−test : $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$

CI for paired t−test : $\bar{x}_d \pm t^* \frac{s_d}{\sqrt{n}}$

CI for $(\mu_A - \mu_B)$: $(\bar{x}_A - \bar{x}_B) \pm t^* \sqrt{\frac{{s_A}^2}{n_A} + \frac{{s_B}^2}{n_B}}$