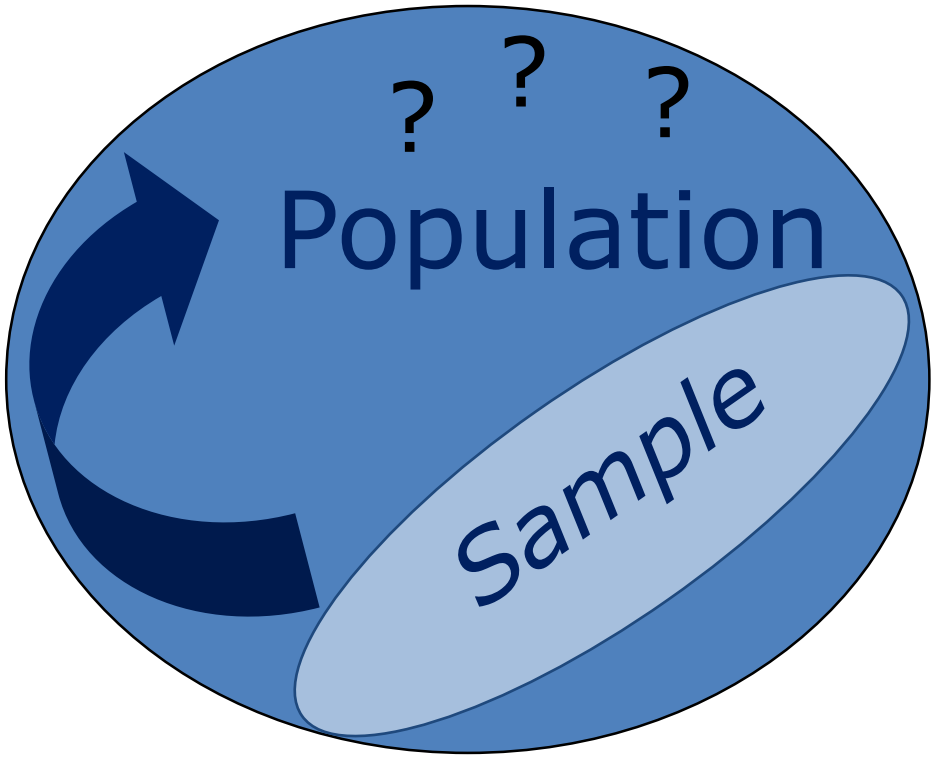
# Lecture 13
# Revision

- Unit Review
- Past Examination Questions

# Answering Research Questions

## Research Questions

We use a SAMPLE to answer questions about a target POPULATION



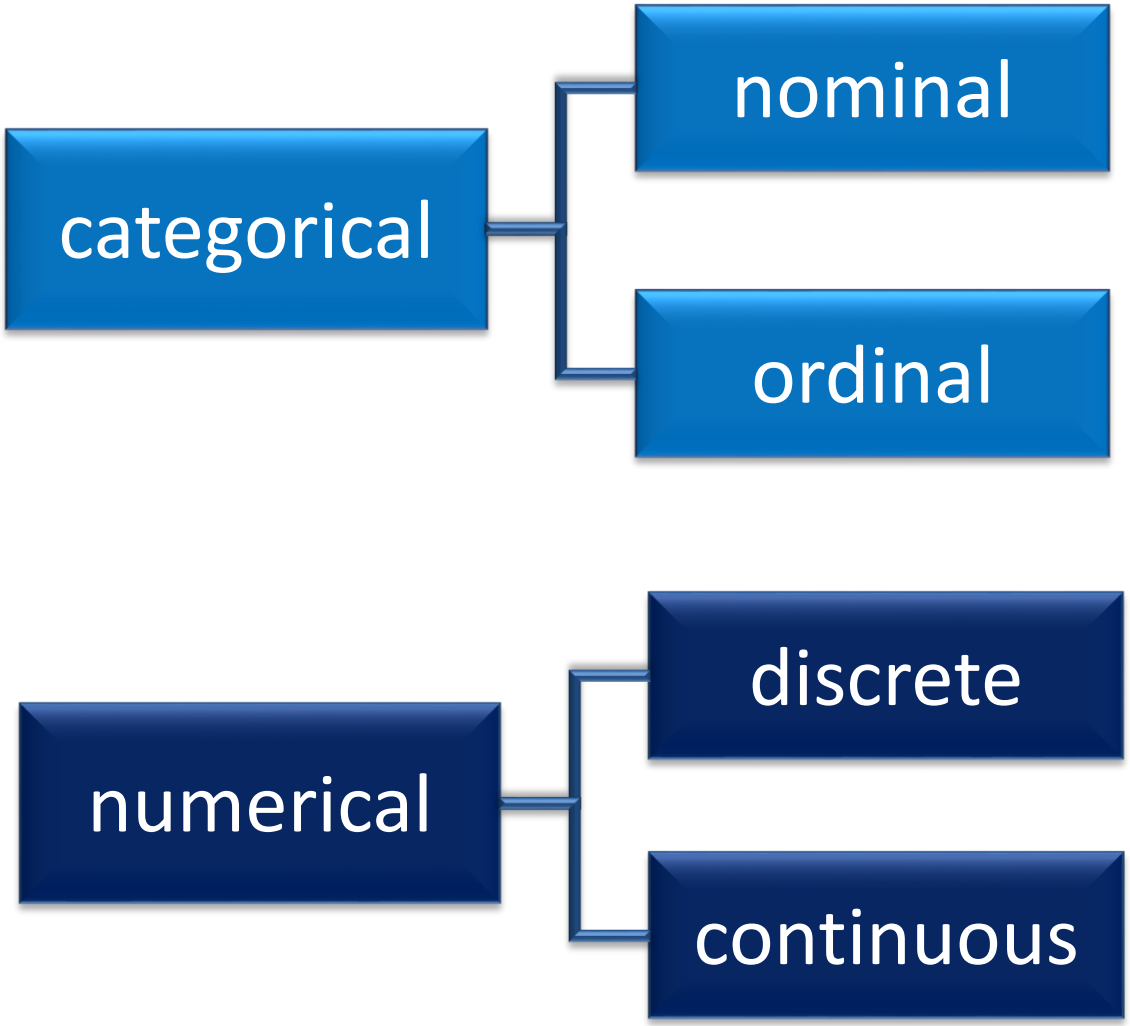
# Notation

| Sample Statistics | estimate | Population Parameters |
|---|---|---|
| Mean | $\bar{y}$ | $\mu$ |
| Median | $\tilde{y}$ | $\tilde{\mu}$ |
| Std.Dev | $s$ | $\sigma$ |
| Variance | $s^2$ | $\sigma^2$ |
| se(mean) | $s/\sqrt{n}$ | $\sigma/\sqrt{n}$ |
| Proportion | $p$ | $\pi$ |
| se(proportion) | $\sqrt{\frac{p(1-p)}{n}}$ | $\sqrt{\frac{\pi(1-\pi)}{n}}$ |
| Least Squares Regression Line | $\hat{y} = a + bx$ | $y = \alpha + \beta x$ |
| Slope | $b$ | $\beta$ |

# Sampling

- Only a *representative sample* should be used to make inferences about the target population. One way to ensure that a sample is representative of the target population is to obtain a *random sample*.
- A *simple random sample*, of a given size n, is one in which *each set* of that size has the same chance of being selected from the target population.
- A *random sample* is one where ***each member*** of the population has the ***same chance*** of being selected.

# Data Classification



# Summarising Data Graphically

| Variable | categorical | numerical | |
|---|---|---|---|
| categorical | clustered bar charts | comparative box plots | bar charts or pie charts |
| numerical | comparative box plots | scatter plots | histograms |
| | bar charts or pie charts | histograms | |

# Summarising Data Numerically

Numerical Variables:

| Descriptive Statistics: Weight of Bilbies | | | | | | |
|---|---|---|---|---|---|---|
| **Variable** | Sex | N | Mean | StDev | Minimum | Maximum |
| **Weight** | Female | 15 | 1.12 | 0.12 | 0.87 | 1.31 |
| | Male | 25 | 1.90 | 0.29 | 1.27 | 2.46 |

Measures of Centre: mean/median

Measures of Spread: range/interquartile range/ standard deviation

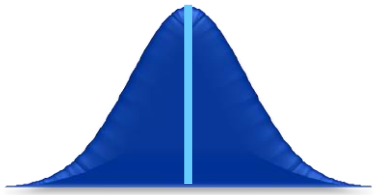
Categorical Variables:

| | Nicotine e-Cigarettes | Nicotine Patches | Total |
|---|---|---|---|
| Quit | 21 | 17 | 38 |
| Did not quit | 268 | 278 | 546 |
| Total | 289 | 295 | 584 |

Tables showing counts/proportions/percentages

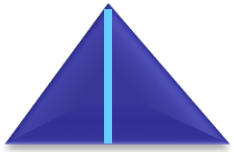
# Population Distributions

*Symmetric*



**Normal**

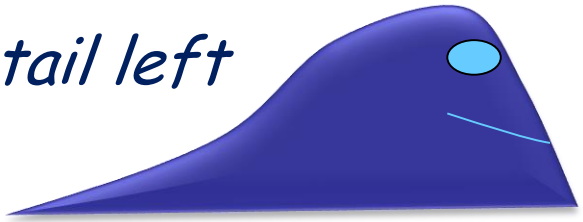
**Uniform**



**Triangular**

*Skewed*

*tail left*



**Skewed Left**

*tail right*

**Skewed Right**

# The normal (z) distribution

If a variable, Y, is drawn from a normal distribution with the mean, $\mu$, known and the standard deviation, $\sigma$, known we can use the normal table to find areas under a normal curve. These areas represent probabilities.

- To find the probability corresponding to a value y, from a normal distribution, calculate its z-score: $z = \frac{y-\mu}{\sigma}$, and look up the corresponding area in the normal (z) table.
- To find a percentile (a y value corresponding to an area under a normal curve), find the corresponding z-score from the normal table and calculate:

$$y = \mu + z \times \sigma$$

# The Central Limit Theorem

The Central Limit Theorem for sample means: in repeated sampling, sample means will follow an approximately normal distribution if the sample size, n, is 'large'. This approximation improves as the sample size increases.

- The standard deviation of sample means (standard error) = $\sigma/\sqrt{n}$
- The z-score for finding a probability for a sample mean, z = $\frac{\bar{y}-\mu}{\sigma/\sqrt{n}}$

The Central Limit Theorem also applies to sample proportions: sample proportions will follow an approximately normal distribution in repeated sampling if the sample size, n is 'large' ie. if both $n\pi$ and $n(1-\pi)$ are ≥ 5.

- The standard deviation of sample proportions (standard error) = $\sqrt{\frac{\pi(1-\pi)}{n}}$
- The z-score for finding a probability for a sample proportion, z = $\frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$

# Confidence Intervals

Confidence Intervals use sample statistics for estimating population parameters with a given level of confidence.

- 95% Confidence Interval for $\mu$ ($\sigma$ known) = $\bar{y} \pm z_{crit} \times {}^{\sigma}/_{\sqrt{n}}$
- 95% Confidence Interval for $\mu$ ($\sigma$ unknown) = $\bar{y} \pm t_{crit} \times {}^{s}/_{\sqrt{n}}$ *(df = n - 1)*

  For 95% intervals: $z_{crit}$ = 1.96, $t_{crit}$ depends on the sample size

- 95% Confidence Interval for $\pi$ = $p \pm z_{crit} \times \sqrt{\frac{p(1-p)}{n}}$
- 95% Confidence Interval for $\mu_d$ = $\bar{y}_d \pm t_{crit} \times {}^{s_d}/_{\sqrt{n_d}}$ *(df = $n_d$ - 1)*
- 95% Confidence Interval for $\mu_1$ - $\mu_2$ = $(\bar{y}_1 - \bar{y}_2) \pm t_{crit} \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

  *(df = $n_1$ + $n_2$ - 2)*

- 95% Confidence Interval for $\beta$ = $b \pm t_{crit} \times se(b)$ *(df = n - 2)*

# Steps in Hypothesis Testing

## Hypothesis Test

**H** ***Hypotheses:***

State the *null* and the *alternative* hypotheses in terms of the *parameter* of interest.

**A** ***Assumptions:***

Check the underlying assumptions of the test.

**T** ***Test Statistic:***

Calculate the test statistic.

**P** ***p-value:***

Obtain the p-value for the test from the distribution of the test statistic.

**D** ***Decision:***

If the p-value is less than 0.05 (the significance level), reject the null hypothesis. If the p-value is not less than 0.05, do not reject the null hypothesis.

**C** ***Conclusion:***

Write a conclusion to the original research question in terms of the target population.

# Formula Sheet 1

## *Statistical Formulas*

$$z = \frac{y - \mu}{\sigma} \qquad z = \frac{\bar{y} - \mu}{\sigma / \sqrt{n}} \text{ (for sample mean)} \qquad z = \frac{p - \pi}{\sqrt{\pi\,(1-\pi)/n}} \text{ (for sample proportion)}$$

One sample z test for population mean: $H_0$: $\mu = \mu_0$ $H_1$: $\mu \neq \mu_0$

$$z = \frac{\bar{y} - \mu_0}{\sigma / \sqrt{n}} \qquad 95\%\ \text{CI:}\ \ \bar{y} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

One sample z test for population proportion: $H_0$: $\pi = \pi_0$ $H_1$: $\pi \neq \pi_0$

$$z = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \qquad 95\%\ \text{CI:}\ \ p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

One sample t test for population mean: $H_0$: $\mu = \mu_0$ $H_1$: $\mu \neq \mu_0$

$$t = \frac{\bar{y} - \mu_0}{s / \sqrt{n}}, \quad (\text{df: } \nu = n-1) \qquad 95\%\ \text{CI:}\ \ \bar{y} \pm t_\nu \frac{s}{\sqrt{n}}$$

# Formula Sheet 2

Paired t test for comparing population means: $H_0: \mu_d = 0$ $H_1: \mu_d \neq 0$

$$t = \frac{\bar{y}_d - \mu_0}{s_d/\sqrt{n}}, \quad (\nu = n_d - 1)$$

$$95\%\ \text{CI:}\ \ \bar{y}_d \pm t_\nu \frac{s_d}{\sqrt{n}}$$

Two sample t test for comparing two population means:
$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, (\nu = n_1 + n_2 - 2)$$

$$95\%\ \text{CI:}\ (\bar{y}_1 - \bar{y}_2) \pm t_\nu s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

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

Linear Regression: t test for population slope: $H_0: \beta = 0$ $H_1: \beta \neq 0$

$$t = \frac{b}{se(b)}, \quad (\nu = n - 2$$

$t_{n-2}$

Chi-square goodness of fit test:
$H_0$: $\pi_1, \pi_1, \ldots, \pi_k$ are as claimed
$H_1$: proportions are not as claimed

$$\chi^2 = \sum_{j=1}^{c} \frac{(O_j - E_j)^2}{E_j} \quad (\nu = c - 1)$$

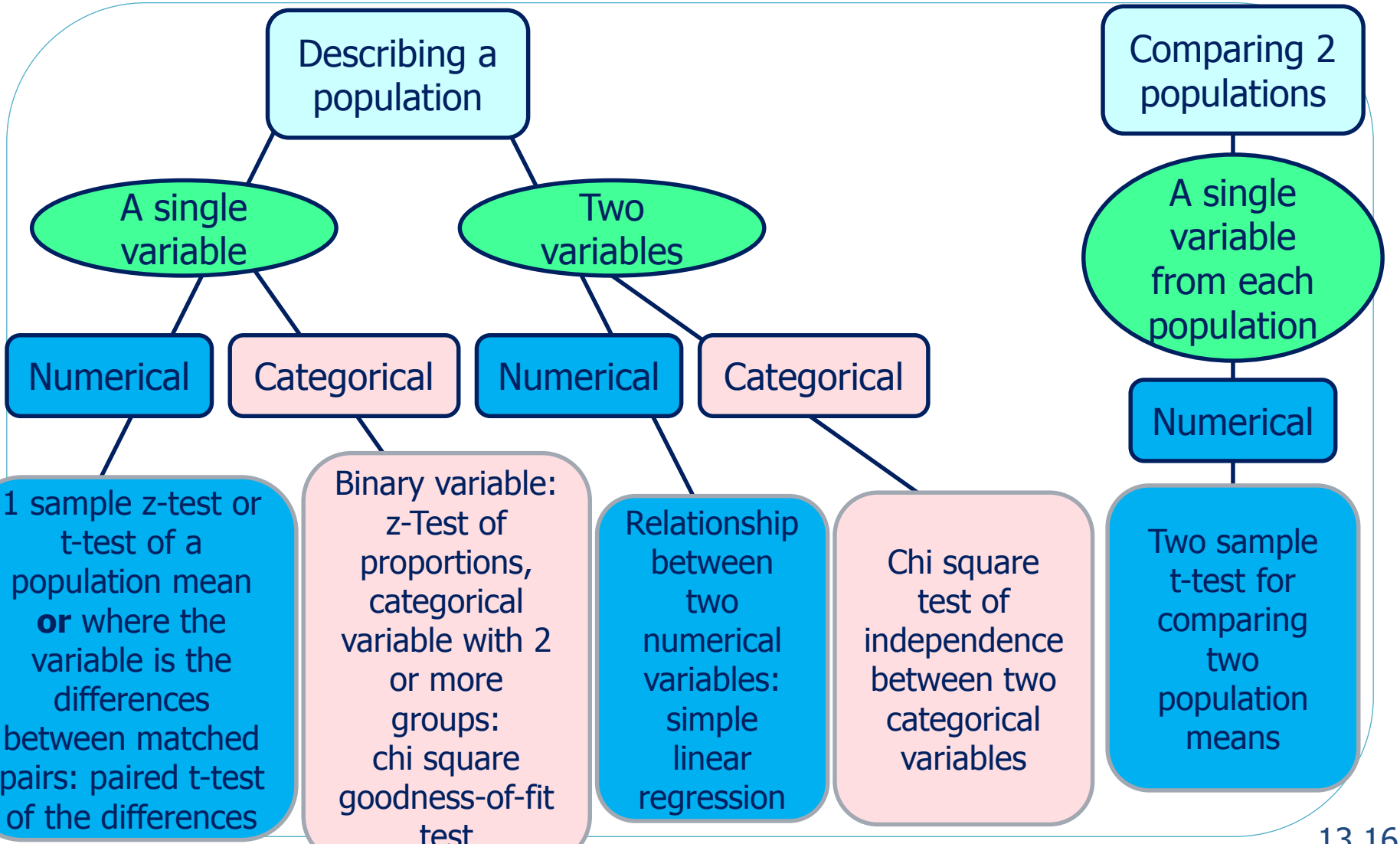
Chi-square test of independence:
$H_0$: There is no association
$H_1$: There is an association

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (\nu = (r-1)(c-1))$$

# Graphs and Hypothesis Tests

| DATA | Categorical | Numerical | |
|---|---|---|---|
| Categorical | ○ Clustered Bar Chart<br>○ Chi Square Test of Independence for the Association between Two Categorical Variables | ○ Comparative Box Plots<br>○ Two Sample t-Test Comparing Two Population Means | ○ Simple Bar Chart<br>○ One Sample z-Test of Proportions/ Chi Square Goodness-of-Fit Test |
| Numerical | | ○ Scatter Plot<br>○ Simple Linear Regression – Test for a Linear Relation Between Two Numerical Variables | ○ Histogram<br>○ One sample z/t-Test for a Population Mean/Paired t-Test for the Differences Between Two Population Means |

# Another Summary



# A Study

A medical trial involved recording the blood pressure and heart rates of a sample of adults who were aged between 18 and 45 years. One of the reasons for the trial was to study how measurements of blood pressure taken in doctors' clinics were related to measurements taken by a device worn by the patient at home, so that future observations could be taken from home measurement. The following information was recorded for the sample of patients at the start of the trial (baseline measurements).

Some of the variables recorded are listed below.

| **Variable** | **Description** |
|---|---|
| Sex | Sex (Male, Female) |
| Smoke | Number of Cigarettes smoked per day (None, 1-5, 6-10, 11-20) |
| BMI | Body Mass Index (100 * weight (kg) / height $(m)^2$ ) |
| Exercise | Amount of usual exercise (None, Light, Moderate) |
| Age | Age (years) |
| SBP_Clinic | baseline Systolic Blood Pressure measured in clinic (mm Hg) |
| DBP_Clinic | baseline Diastolic Blood Pressure measured in clinic (mm Hg) |
| HR_Clinic | baseline Heart Rate measured in clinic (beats per minute) |
| SBP_Home | baseline Systolic Blood Pressure measured at home (mm Hg) |
| DBP_Home | baseline Diastolic Blood Pressure measured at home (mm Hg) |
| HR_Home | baseline Heart Rate measured at home (beats per minute) |

*Source OzDasl and Schork, M. A., and Remington, R. D. (2000). Statistics with Applications to the Biological and Health Sciences 3rd Edition. Prentice-Hall (adapted)*

# Question 1

**Question 1 (7 marks)**

a. If possible, list one variable which is:

Numerical

Ordinal

Nominal

b. Would it be possible for the variable 'Age' to be recorded as an ordinal variable? How would the observations be recorded in this case?

# Question 1

c. If all the information had been recorded from patients of one doctor, describe an appropriate target population for this trial

d. What would be an appropriate graph to use to investigate whether there is an association between the amount of usual exercise (none/light/medium) and the number of cigarettes smoked per day (none/1-5/6-10/11-20) among adults?

e. What would be an appropriate graph to use to investigate whether there is a difference between average systolic blood pressure for males and females?

# Question 1

c. If all the information had been recorded from patients of one doctor, describe an appropriate target population for this trial

**Adult patients of this doctor who are aged from 18 to 45**

d. What would be an appropriate graph to use to investigate whether there is an association between the amount of usual exercise (none/light/medium) and the number of cigarettes smoked per day (none/1-5/6-10/11-20) among adults?

**Clustered bar chart**

e. What would be an appropriate graph to use to investigate whether there is a difference between average systolic blood pressure for males and females?

**Comparative box plots**

# Information for Questions 2 and 3

**Information for Question 2 and Question 3**

The following table summarises the data from a sample of 200 patients in this trial for the variables Smoke and Sex. Use the data in this table in answering Questions 2 and 3.

**Tabulated statistics: Sex, Smoke**

Rows: Sex Columns: Smoke

| | None | 1-5 | 6-10 | 11-20 | All |
|---|---|---|---|---|---|
| Female | 47 | 5 | 0 | 2 | 54 |
| Male | 111 | 10 | 9 | 16 | 146 |
| All | 158 | 15 | 9 | 18 | 200 |

Cell Contents: Count

# Question 2

a. What proportion of females in the sample do not smoke?

b. Calculate a 95% confidence interval to estimate the true proportion of females in the target population who do not smoke.

# Question 2

c. **Research Question:** Are males and females represented in equal proportions in this trial? Carry out an appropriate hypothesis test to address this research question.

# Question 3

**Research Question:** Do adults smoke in the proportions claimed by a researcher? A researcher claims that 75% of people are non-smokers, 15% smoke 1-5 cigarettes per day, and the remainder are split equally between smoking 6-10 and 11-20 cigarettes per day. Use the summary table on slide 13.18 to carry out an appropriate hypothesis test to address the research question.

# Question 4

**Research Question:** Do males and females have similar patterns of exercise? The following Minitab output was generated from a sample of patients in the trial.

**Tabulated statistics: Sex, Exercise**

Rows: Sex  Columns: Exercise

| | None | Light | Moderate | All |
|---|---|---|---|---|
| Female | 41 | 7 | 5 | 53 |
| | 34.07 | 8.11 | 10.82 | |
| | 1.4090 | *** C *** | 3.1276 | |
| Male | 85 | 23 | 35 | 143 |
| | *** B *** | 21.89 | 29.18 | |
| | 0.5222 | 0.0565 | 1.1592 | |
| All | 126 | 30 | 40 | 196 |

Cell Contents: Count
Expected count
Contribution to Chi-square

Pearson Chi-Square = * , DF = * , P-Value = 0.040

a. One of the expected counts in this output on the previous slide has been marked '*** B ***'. Calculate this value.

b. One of the standardised discrepancies has been marked '*** C ***' on the output above. Calculate this value.

# Question 4

c. Use an appropriate hypothesis test to determine whether there is an association between sex and level of exercise.

# Information for Questions 5 and 6

Recall that one of the reasons for the trial was to study how measurements of blood pressure taken in a doctor's clinic were related to measurements taken by a device worn by the patient at home, so that future observations could be taken from home measurement. A small sample (n=30) of patients had measurements of blood pressure (systolic and diastolic) and heart rate repeated at home 3 months after the baseline measurements were taken. The variables of interest in these questions are:

| | |
|---|---|
| SBP_Home | baseline Systolic Blood Pressure measured at home (mm Hg) |
| DBP_Home | baseline Diastolic Blood Pressure measured at home (mm Hg) |
| HR_Home | baseline Heart Rate measured at home (beats per minute) |
| SBP_Home_3 | Systolic Blood Pressure measured at home 3 months later (mm Hg) |
| DBP_Home_3 | Diastolic Blood Pressure measured at home 3 months later (mm Hg) |
| HR_Home_3 | Heart Rate measured at home 3 months later (beats per minute) |

# Question 5

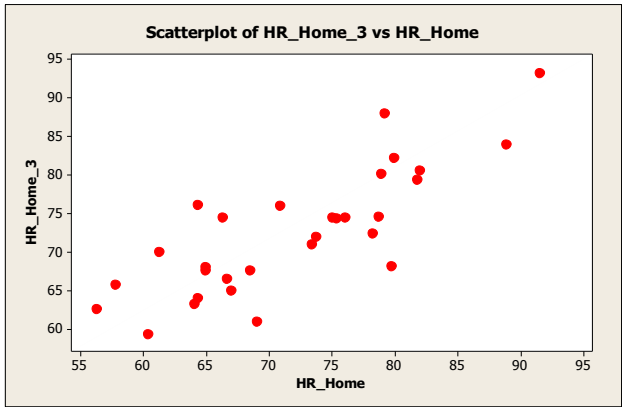
**Research Question:** Is baseline heart rate measured at home a good predictor of heart rate measured at home 3 months later?

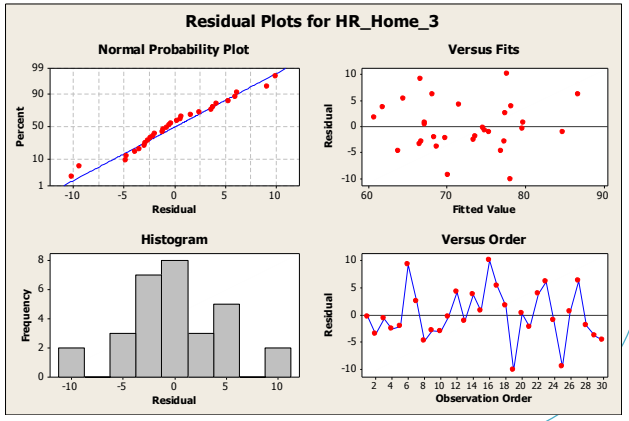
**Regression Analysis: HR_Home_3 versus HR_Home**

The regression equation is
HR_Home_3 = 19.1 + 0.740 HR_Home

| Predictor | Coef | SE Coef | T | P |
|---|---|---|---|---|
| Constant | 19.058 | 7.316 | 2.60 | 0.015 |
| HR_Home | 0.7399 | 0.1008 | * | * |

S = 4.82545  R-Sq = 65.8%



# Question 5

a. Comment on the relation between baseline heart rate and heart rate 3 months later.

b. Did the person with the highest baseline heart rate also have the highest heart rate 3 months later? What is the approximate heart rate after 3 months of person with the highest baseline heart rate?

c. Indicate the value of the correlation between the heart rate at 3 months and the baseline heart rate and comment on this value.

# Question 5

d. Carry out an appropriate hypothesis test to determine whether there is a significant relationship between heart rates at 3 months and baseline heart rates (n = 30).

# Question 6

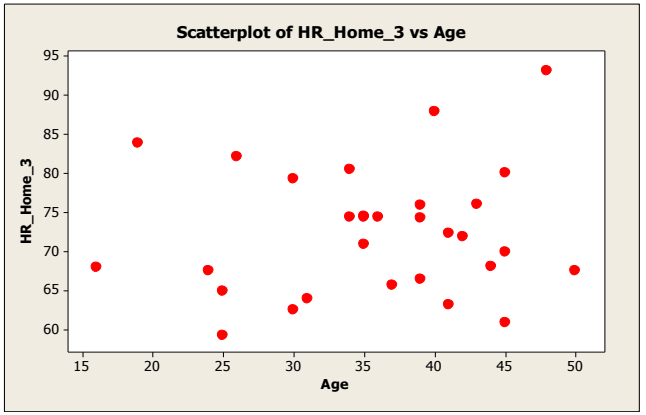
**Research Question:** Is heart rate measured after 3 months related to a patient's age?

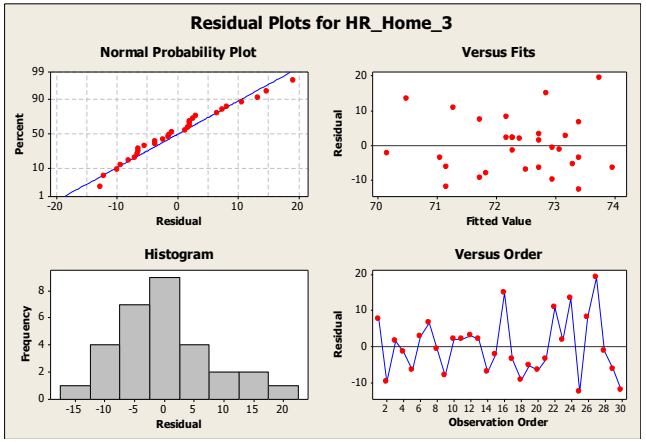
**Regression Analysis: HR_Home_3 versus Age**

The regression equation is
HR_Home_3 = 68.4 + 0.112 Age

| Predictor | Coef | SE Coef | T | P |
|---|---|---|---|---|
| Constant | 68.367 | 6.558 | 10.42 | 0.000 |
| Age | 0.1122 | 0.1785 | 0.63 | 0.535 |

S = 8.19525  R-Sq = 1.4%



# Question 6

a. Interpret the goodness of fit statistic for the relationship between heart rate at 3 months and the age of the patient.

b. Calculate a 95% confidence interval for the population slope of the line which relates heart rate at 3 months and the age of the patient. Comment on this confidence interval. Recall n = 30.

# Question 6

**Use all of your results from both Questions 5 and 6 to answer the rest of this question.**

c. If possible, calculate the predicted heart rate after 3 months for a patient whose baseline heart rate was 65 beats per minute. If it is not possible to give this prediction, explain why not.

d. If possible, calculate the predicted heart rate after 3 months for a patient who is 30 years old. If it is not possible to give this prediction, explain why not.

e. Which is the better predictor of heart rate after 3 months – baseline heart rate or age?

# Another Study

How difficult is it to maintain your balance while concentrating? Is it more difficult when you are older? Eight elderly people and eight young people were subjects in an experiment. Each subject stood barefoot on a "force platform" and was asked to maintain a stable upright position and to react as quickly as possible to an unpredictable noise by pressing a hand held button. The noise came randomly and the subject concentrated on reacting as quickly as possible. The variable FBSway indicates how far (in mm) each subject swayed in the forward/backward direction. The variable SideSway indicates how far (in mm) each subject swayed in the side to side direction.

*Source: OzDasl and Teasdale, N., Bard, C., La Rue, J., and Fleury, M. (1993). On the cognitive penetrability of posture control. Experimental Aging Research (adapted)*

# Minitab Output for Questions 7 and 8

**Two-Sample T-Test and CI: FBSway, SideSway**

```
Two-sample T for FBSway vs SideSway

            N   Mean  StDev  SE Mean
FBSway     16  20.75   4.91      1.2
SideSway   16  17.75   7.37      1.8

Difference = mu (FBSway) - mu (SideSway)
Estimate for difference:  3.00
95% CI for difference:  (-1.52, 7.52)
T-Test of difference = 0 (vs ≠): T-Value =  *  P-Value = 0.185  DF = 30
Both use Pooled StDev = 6.2610
```

**Paired T-Test and CI: FBSway, SideSway**

```
Paired T for FBSway - SideSway

              N   Mean  StDev  SE Mean
FBSway       16  20.75   4.91     1.23
SideSway     16  17.75   7.37     1.84
Difference   16   3.00   6.31     1.58

95% CI for mean difference: (-0.36, 6.36)
T-Test of mean difference = 0 (vs not = 0): T-Value =  *  P-Value = 0.077
```

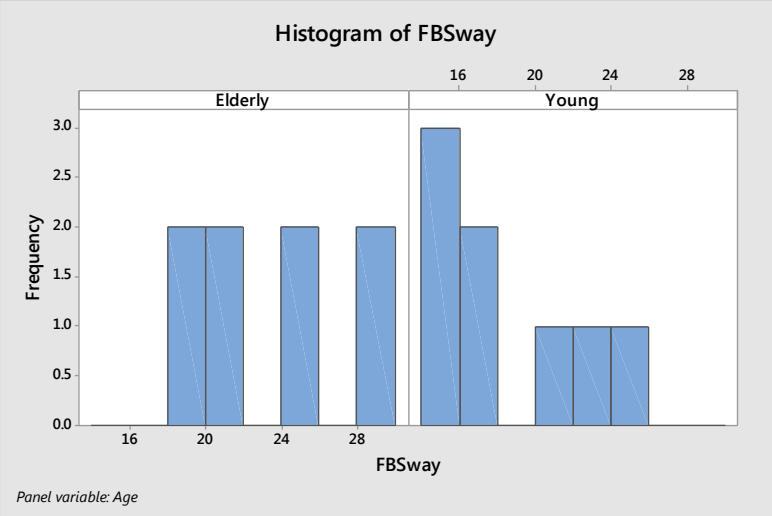
**Paired T-Test and CI: FBSway_Elderly, FBSway_Young**

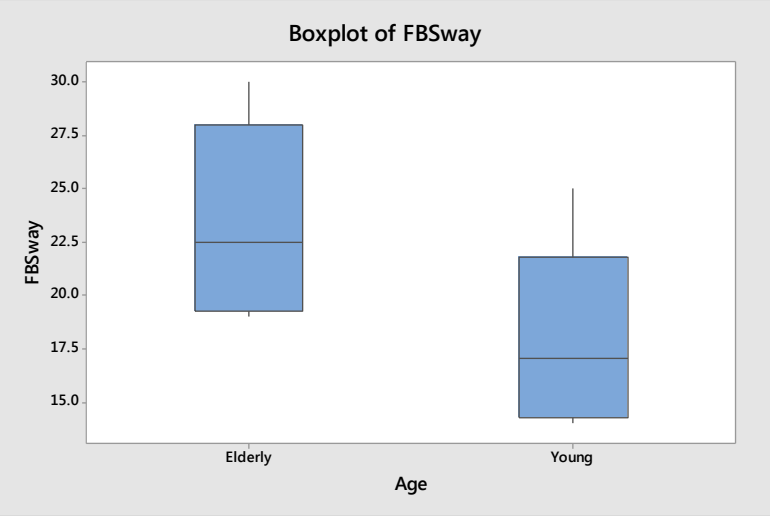
```
Paired T for FBSway_Elderly - FBSway_Young

                 N   Mean  StDev  SE Mean
FBSway_Elderly   8  23.38   4.37     1.55
FBSway_Young     8  18.13   4.09     1.44
Difference       8   5.25   6.73     2.38

95% CI for mean difference: (-0.38, 10.88)
T-Test of mean difference = 0 (vs not = 0): T-Value =  *  P-Value = 0.063
```

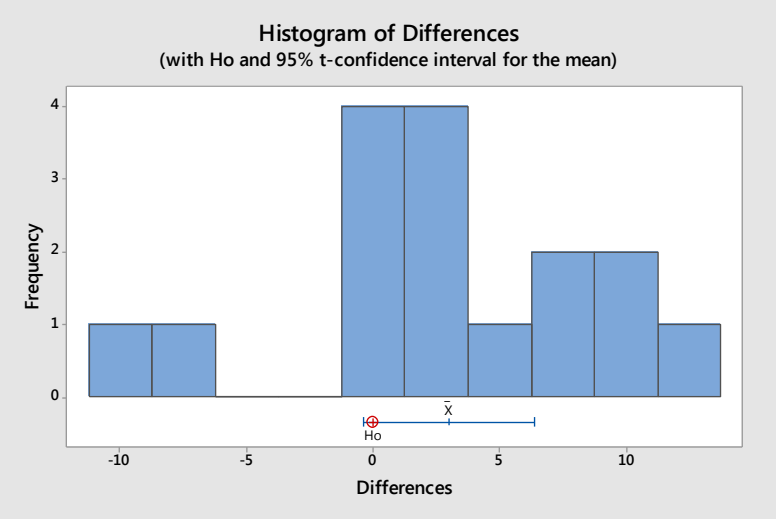
**Two-Sample T-Test and CI: FBSway_Elderly, FBSway_Young**

```
Two-sample T for FBSway_Elderly vs FBSway_Young

                 N   Mean  StDev  SE Mean
FBSway_Elderly   8  23.38   4.37      1.5
FBSway_Young     8  18.13   4.09      1.4

Difference = mu (FBSway_Elderly) - mu (FBSway_Young)
Estimate for difference:  5.25
95% CI for difference:  (0.71, 9.79)
T-Test of difference = 0 (vs ≠): T-Value =  *  P-Value = 0.026  DF = 14
Both use Pooled StDev = 4.2321
```

# More Output for Questions 7 and 8



# Question 7

**Research Question:** Does the distance people sway in the forward/backward direction differ for older people and younger people? Use the output provided to carry out an appropriate hypothesis test to answer the research question:

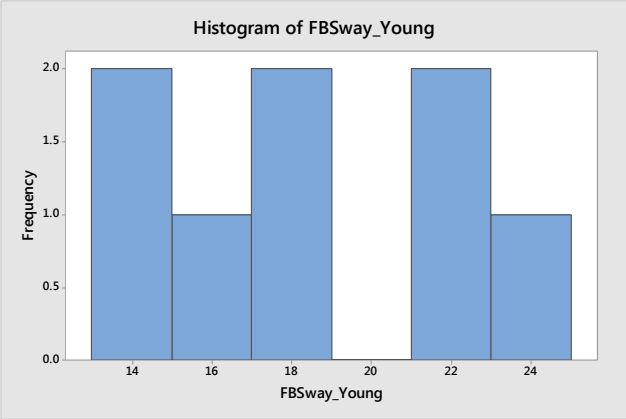
# Question 8

a. **Research Question:** Regardless of age, do people sway the same distance in the forward/backward direction and the side-to-side direction?

Use appropriate parts of the output provided to carry out an appropriate hypothesis test answer the research question:

# Question 8

**b. Research Question:** Do young people sway an average of 14mm in a forward/backward direction?




**Descriptive Statistics: FBSway_Young**

| Variable | N | Mean | StDev | Minimum | Maximum |
|---|---|---|---|---|---|
| FBSway_Young | 8 | 18.13 | 4.09 | 14.00 | 25.00 |

Use appropriate parts of the output provided to carry out an appropriate hypothesis test answer the research question.

# Question 8

b. Hypothesis Test:

# Question 9

Recall from lectures that adult IQ scores are known to be normally distributed, with a mean of 100 and a standard deviation of 15.

a. What is the probability that a random sample of 25 adults would have an average IQ of between 105 and 110?

# Question 9

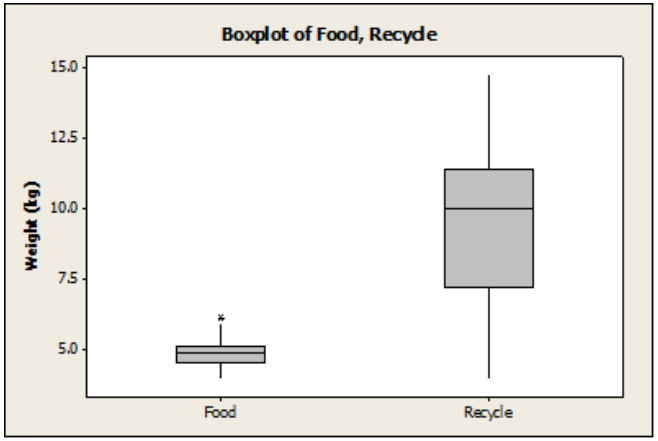
b. A research group wants to recruit adults of moderately high intelligence (defined as IQ greater than 135) for a study. They can select their sample from 1000 adults who have agreed to participate. How many adults would they expect to meet their criteria?

# Question 10

a. Recent research has suggested that the true proportion of households in a particular area which have soil with a high level of lead contamination is 0.20. If we take a random sample of 50 households, what is the probability that between 8 and 15 of them will have soil with a high level of lead contamination.

# Question 10

b. The boxplot below displays the weights of food waste and recycling materials in a garbage collection for a sample of households. Indicate whether each of the statements below is likely to be true or false.




| Statement | True or False? |
|---|---|
| The weight of food waste is typically higher than the weight of recycling materials | |
| The weight of food waste is less variable than the weight of recycling materials | |
| The weight of recycling materials is typically higher than the weight of food waste | |
| One household has an unusually low weight of food waste | |