More practice on choosing which statistical test to use:

1. Bogton Council decide to see whether performance-related pay would improve morale amongst their lavatory cleaners. Each month, twenty lavatory cleaners are paid on the basis of the length of the bristles on their lavatory brush (on the assumption that the harder they have worked, the shorter their bristles will be). Another twenty are paid their usual near-subsistence-level wages, regardless of how hard they work. After 6 months, each worker is asked to rate how happy they are in their job, using a seven-point scale. Which test would you use to see if performance-related pay has affected workers' morale?

2. An experimenter wants to know whether experience affects how well shop-keepers can identify children who ask for cigarettes but are under the legal age for purchasing them. Each of 30 tobacconists is shown a random sequence of 40 photographs of young faces, and asked to decide whether each face is younger or older than the legal age for buying cigarettes. (Half of the faces are aged above the legal age, and half below). The experimenter records the number of correct decisions per participant, and also asks each shop-keeper how long they have been selling cigarettes. (These latter data turn out to be heavily skewed). Which test should the experimenter use to decide whether experience leads to better age-estimation in this group?

3. It's often said that you're hungry again soon after a Chinese meal. An experimenter puts this to the test. There are four conditions, and each participant does each one, on a different day of the week (order of conditions is counterbalanced across participants). In the first, participants eat an Indian takeaway; in the second, they eat a pizza; in the third, they eat a Chinese takeaway; and in the fourth, they eat a Kentyukky Flayed Chicken takeaway. All the meals are equated for bulk of contents and calorific value. The dependent variable is the loudness of each participants' stomach rumblings (in decibels), measured one hour after they have eaten the meal. These measurements are normally distributed, but much more variable for the "KFC" condition than the others. Which test should be used to decide whether there is a difference between these meals in terms of how quickly people get hungry again after eating them?

4. Some TV viewers complain to the BBC that Jeremy Clarkson's programme "Top Gear" is a bad influence on young drivers, given that it extols the virtues of laddishness, speeding and high performance cars. To determine whether there is any foundation to these claims, a researcher uses a speed camera to measure the speeds of 400 drivers on an A-road, the morning before the programme is transmitted. He follows this procedure again, the morning afterwards. Each car is photographed, so that the experimenter can select only those drivers who travelled that route on both occasions, and hence whose speeds were measured twice. The experimenter subtracts each driver's first speed reading from their second, to get a "difference score": a positive score means a driver drove faster on the second occasion, and a negative score means they drove more slowly. The selected drivers were then contacted and asked whether or not they had watched "Top Gear" that week. Which test would you use to see whether drivers who watched "Top Gear" drove faster the following morning than drivers who did not watch it?

5. A researcher is interested in factors affecting reproductive success in *Homo canarywharfensis*, an obscure species of proto-human that inhabits high-altitude habitats in a region of south-east London. Once she has acclimatised them to her presence, she traps a hundred of the males and records the price of their suits. She then releases them back into the wild and follows them for a fortnight, recording how many females each one mates with. Is there a relationship between wealth (as reflected in suit price) and reproductive success (as reflected in how many females each male mates with?) The data for reproductive success are heavily skewed, since most of the males attract no females.

6. The local Sussex ale, Harvey's Best bitter, is reputed to be imbued with truly magical medicinal properties, as well as having an especially delicious flavour, a unique golden colour and a beautiful yeasty head. To investigate its effects, a researcher asks four groups of cyclists to cycle up Ditchling Beacon (the highest point on the South Downs). One group drink no Harvey's beforehand; another group drink one pint of Best each; a third group drink two pints each; and a fourth group drink four pints each. The dependent variable is how fast each cyclist gets from the bottom of the Beacon to the top. Which test would you use to see if drinking Harvey's affects the cyclists' speed of ascent?

7. It is said that every time someone prints off an email, a penguin dies. To put this to the test, a researcher flies to the South Pole and repeatedly counts the number of penguins, as her colleague at Sussex prints out his emails one at a time. Which test would you use to see if there is a relationship between printing off emails and penguin mortality?

8. A researcher investigates four different methods for coping with extreme stress. Each person attempts to assemble an IKEA flat-pack wardrobe (the stress-induction phase of the study), and is then allocated randomly to one of four groups. Those in the first group practise yoga for twenty minutes; those in the second group engage in deep breathing for a similar amount of time; those in the third group spend twenty minutes in a Harvey's pub, drinking Best bitter; and those in the fourth group simply scream at the top of their voice for twenty minutes. Each participant then provides a rating on a 0-10 scale of how stressed they feel. Which test would you use to determine whether the four methods differ in their effectiveness for relieving stress?

9. To determine whether young chldren find "Dr. Who" scary, a researcher asks the parents of thirty six-year olds to keep a record of how many nightmares each child has on Saturday night (after watching "Dr. Who") and Sunday night (after watching "Songs of Praise"). Which test would you use to see if watching "Dr. Who" is associated with more nightmares than watching "Songs of Praise"?

10. To determine whether young chldren find "Dr. Who" scary, a researcher asks the parents of thirty six-year olds to rate how frightened they think their child is on Saturday night (after watching "Dr. Who") and Sunday night (after watching "Songs of Praise"). Which test would you use to see if parents think their children are more frightened by watching "Dr. Who" than by watching "Songs of Praise"?

11. 200 men and 150 women are asked to decide which one of the following features is most important to them when they choose a new car: price, performance, safety level, roominess, or colour. Which test would you use to see if men and women differ in their preferences?

12. An experimenter investigates the accuracy of fortune-tellers' predictions. She asks each of fifteen fortune-tellers, and each of twenty students, to make ten specific predictions about what will happen to her in the next month. She then records, for each of these people, how many of these predictions come true. Which test should she use to see if the fortune-tellers are more accurate in their predictions than the students?

13. The experimenter from the previous study returns to each of the participants and tells them that none of their predictions came true. She then asks each of the participants to estimate their level of psychic ability on a seven-point scale. Which test should the experimenter use to determine whether this negative feedback about their performance affects the fortune-tellers and students differently?

14. A study looks at the effectiveness of TV adverts in relation to their position in the ad-break between programmes. There are three conditions. All participants see the same advert, for "Churn Flakes", but for one group the advert comes at the start of the ad-break; for a second group, it comes in the middle; and for the third group it comes at the end, just before the next program begins. A week later, each participant returns to the lab and sees a sequence of photographs of breakfast cereal boxes, including the box for "Churn Flakes". Their task is to rate each cereal in terms of how much they like it, using a seven-point scale.

15. While sales of traditional classical music CD's are falling, "cross-over" classical performers who sacrifice their integrity for money by producing populist versions of tunes like "Nessun Dorma" are big business. The CD sales of twenty opera singers are examined: ten of these singers are rated as "ugly" by a panel of independent judges, and twenty are rated as "highly attractive". Is the success of these perfomers related to their physical attractiveness?

Answers:

1. This is an independent-measures design, with two conditions, those workers who have performance-related pay and those who don't. We are looking to see if there is a *difference* between these two conditions. The data are ratings of job satisfaction, and so a non-parametric test is called for, the **Mann-Whitney test**.

3. This is a repeated-measures design, with four conditions. Each participant provides a single score for each condition, namely loudness of stomach-rumbling. These are ratio data, but they violate one of the requirements for using a parametric test: there appears to be inhomogeneity of variance, since scores are much more spread out in one condition than the others. The appropriate test is therefore a nonparametric test for differences between three or more conditions within a repeated-measures design - **Friedman's test**.

2. The experimenter has two measurements for each participant (age-estimation accuracy and years of experience) and is interested in the *relationship* between them. It is in essence a correlational design, since the experimenter has no control over either of the independent variables concerned. The data on experience are skewed, and hence not normally distributed; consequently the appropriate test is the nonparametric correlation test, **Spearman's rho**.

4. There are two conditions: drivers who watched "Top Gear", and drivers who didn't. Each participant provides a single score, the difference between how fast they drove on two occasions. These are ratio data: there is a true zero to the scale (someone who drove at the same speed on both occasions produces a difference score of zero), and a difference score of 20 miles per hour is twice a score of 10 miles an hour, half a score of 40 mph, and so on). Assume the data are normally distributed and that there is homogeneity of variance, since the question gives no indications to the contrary. We are looking for a difference between two conditions; it's an independent-measures design; and the data appear to satisfy the requirements for a parametric test. The appropriate test is therefore an **independent-measures t-test**.

5. The researcher is interested in the *relationship* between two variables, wealth and reproductive success. Each participant provides a score on each of these variables. These data are ratio measurements (price, and number of partners), but the data on one of the variables are skewed, and hence not normally distributed. Therefore the appropriate test is a non-parametric correlation test, **Spearman's rho**.

6. We have an independent-measures design, with four conditions (varying in the amount of beer that was drunk). Each participant belongs to one condition only, and provides a single score. Speed of ascent is a ratio measure, and we are given no indications that the data do not satisfy the other requirements for a parametric test. Therefore the appropriate test is a **one-way independent-measures ANOVA.**

7. We are interested in whether there is a linear relationship between the number of penguins and the number of emails printed off. It happens to be a causal relationship (email printing causes penguin death), but this is essentially a correlational design. (We are looking for a negative correlation: as number of printed emails goes up, so the number of breathing penguins goes down). Both variables (number of emails and number of penguins) are measurements on ratio scales, and we are given no reason to suppose that the data do not satisfy the other requirements for a parametric test. The appropriate test is therefore **Pearson's r**.

8. We have an independent-measures design: there are four different conditions, and each participant takes part in only one of them. Each participant provides a single score, their rated stress level. Rating scores are ordinal data. Consequently we want an independent-measures, non-parametric test that compares performance in three or more conditions - the **Kruskal-Wallis test**.

9. This is a repeated-measures design, since each child participates in both conditions of the study. Each child provides two scores: the number of nightmares following watching "Dr. Who", and the number of nightmares after watching "Songs of Praise", and we are looking for a difference between these two conditions. "Number of nightmares" is a measurement on a ratio scale. We can assume that the data are normally-distributed and that there is homogeneity of variance, since we are given no indications to the contrary. A parametric test can therefore be used. The appropriate test is therefore the **repeated-measures t-test.**

10. This is a repeated-measures design, since each child participates in both conditions of the study. Each parent provides two scores for their child: a rating of how frightened they think their child was after watching "Dr. Who", and a rating of how frightened their child was after watching "Songs of Praise". We are looking for a difference between these two conditions. We can assume that the data are normally-distributed and that there is homogeneity of variance, since we are given no indications to the contrary. However, ratings are measurements on an ordinal scale, and so a nonparametric test is called for. The appropriate test is therefore the nonparametric equivalent of a repeated-measures t-test - the **Wilcoxon test.**

11. These are frequency data. We have a 2 x 5 contingency table. All we know is how many people fall into each of ten categories, the various permutations of gender (male or female) and car attribute (price, performance, safety level, roominess, or colour). Thus we have the number of men who think price is most important, the number of women who think performance is most important, the number of men who think colour is most important, etc. We are interested in whether there is an association between the two variables of gender and car attribute: is the pattern of preferences non-random, and is it different for men and women? The appropriate test is therefore the **Chi-Square test of association**.

12. This is an independent-measures design, with two groups: fortune-tellers and students. Each participant provides one score, the number of predictions that came true.

Accuracy data like these are measurements on a ratio scale. We can assume that the data are normally-distributed and that there is homogeneity of variance, since we are given no indications to the contrary. Since the data satisfy the requirements for a parametric test, we can use the **independent-measures t-test.**

13. This is an independent-measures design, with two groups: fortune-tellers and students. We are looking for a difference between these two groups, in terms of the effects of negative feedback on them. Each participant provides one score, their rating of their own psychic ability after being told they are rubbish at predicting the future. We can assume that the data are normally-distributed and that there is homogeneity of variance, since we are given no indications to the contrary. However, rating data like these are measurements on an ordinal scale, and so a non-parametric test is called for. The appropriate test is therefore the non-parametric equivalent of an independent-measures t-test - the **Mann-Whitney test.**

14. This is an independent-measures design, with three groups: those who saw the advert at the start of the ad-break, those who saw it in the middle, and those who saw it at the end. Each participant provides one score: their rating of how much they like "Churn Flakes" a week later. These are ordinal data (ratings) and so a non-parametric test is called for. The appropriate non-parametric test for an independent-measures design with three or more conditions is the **Kruskal-Wallis test.**

15. This is an independent-measures design with two conditions: "ugly" singers and "attractive" singers. The dependent variable is the number of CD's sold by each singer. Thus we have a score per participant; each participant falls into one of two conditions (ugly or attractive); and we are looking to see if there is a difference between the two conditions. "Number of CD's sold" is a ratio measure. We can assume that the data are normally distributed, and that there is homogeneity of variance, since there are no indications to the contrary. The appropriate test is the **independent-measures t-test.**