Write the name of an appropriate statistical procedure for the following situations – for instance "Wilcoxon–Mann–Whitney test" or "chi-squared for goodness of fit" or "t-test for difference in means".

A. We have measured amount of body fat in random samples of male and female elephant seals from three different breeding grounds. Does the mean amount of body fat differ across breeding grounds, after accounting for differences in the sexes?

F-test for two-way ANOVA; just "ANOVA" almost correct

B. We have measured people's heights in a random sample; do men and women differ in (population) average height?

two–sample t-test, or Wilcoxon–Mann–Whitney

C. We have recorded incomes in a random sample of 10 female and 10 male Angelenos; do men and women differ in (LA population) average income? The distribution of incomes is strongly skewed to the right (i.e. has a few values much larger than the others).

Wilcoxon–Mann–Whitney; partial credit for two–sample t-test

D. In a random sample of tumors we have counted the number that fall into three histological grades, which we previously predicted should occur in the ratio 4:2:1. Are our data consistent with this?

chi-squared (for single-sample proportions)

Here are several statements from academic studies. What more would you like to know to judge the strength of the statistical evidence and the importance of the results?

- "We pit luna moths against big brown bats [...]. Moth tails lured bat attacks to these wing regions during 55% of interactions between bats and intact luna moths."

  We know the result (55%), but there is no estimate of uncertainty: a confidence interval for the result or at least a sample size would be important.

- "We observed significant improvement in memory task performance under drug treatment relative to placebo in the aMCI cohorts at the 62.5 and 125 mg BID doses of levetiracetam."

  We know the improvement is "signficant" – but, that probably means "statistically significant", so: I'd like to know how big was the improvement, in interpretable real-world terms (percent improvement, or an effect size).

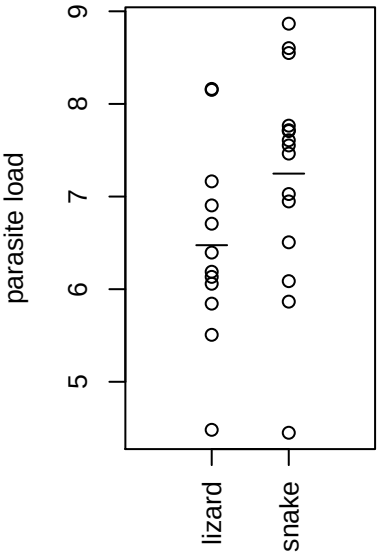
- "For each root, the force at the time of fracture was recorded in Newtons. The fracture values of three experimental and negative control groups were significantly higher than the positive control group."

  Again, "significantly higher" means "statistically" – how much higher was the force, on average?

- "The study also found that young adults with schizophrenia who abused cannabis as teens performed about 26 percent more poorly on memory tests than young adults with schizophrenia who never abused cannabis."

  Two issues here: First, what is the level of certainty about that 26% – we need a confidence interval or P-value. Second, how did they control for confounding factors, if at all?

We are interested in whether a certain parasite prefers some species over others, and so have measured parasite load (mg of parasite per animal) in random samples of 12 fence lizards and 15 garter snakes in the same habitat, finding that the average load in lizards is 6.474, with an SD of 1.047, and the average load in snakes is 7.247, with an SD of 1.161. The data are shown on the right.



A. Estimate the effect size of the difference in average load, assuming the population SDs are equal to 1.0 for both lizards and snakes, and briefly say what it means, in words.

Effect size is

$$\frac{7.247 - 6.474}{1.0} = 0.773,$$

i.e. the difference between the samples is about 3/4th of the SD of the variation within samples.

B. State an appropriate null and alternative hypothesis for a two–sample t test of the difference in means.

$H_0 : \mu_{\text{lizards}} = \mu_{\text{snakes}}$; population mean load is the same for each

$H_A : \mu_{\text{lizards}} \neq \mu_{\text{snakes}}$; population mean load differs

C. Carry out the t test of these hypotheses, with $\alpha = 0.05$.

$$\mathsf{SE}_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{1.047}{12} + \frac{1.161}{15}} = 0.4057709$$

$$df = \frac{(1.047^2 + 1.161^2)^2}{1.047^4/11 + 1.161^4/14} = 24.99266$$

$$t_s = \frac{0.773}{0.4057709} = 1.905016$$

$$.06 < P < .08$$

D. What can we conclude about whether the parasite prefers snakes or lizards?

We have weak evidence that there is a difference in parasite load, but would need a larger sample size to find out more confidently. If the difference is real, then based on these data it is probably fairly substantial (snakes have about 10% more parasites than lizards, effect size of 0.7); but it is an observational study, so the difference might be due to factors other than parasite preference.

We hypothesize that students feel better on Mondays than on Fridays, thanks to having more time for sleep. A randomly chosen group of 40 USC students reported their general mood on a Monday and the following Friday. Of these, 26 reported having a better mood on Monday, and 14 reported having a better mood on Friday.

A. State the null and alternative hypotheses, in words and then in symbols.

H_0 : $\mathbb{P}\{\text{mood better on Monday}\} = \mathbb{P}\{\text{mood better on Friday}\}$,
students were just as likely to be in a better mood on Monday or on Friday

H_A : $\mathbb{P}\{\text{mood better on Monday}\} > \mathbb{P}\{\text{mood better on Friday}\}$,
students were more likely to be in a better mood on Mondays.

B. Test the null hypothesis using a sign test, and briefly state your conclusion.

$N_+ = 26$, $N_- = 14$, $B_s = 26$, $n_d = 40$; since this is beyond Table 7 in the book, we need to use a computer to get: $P = 0.04034523$. *Note: on the final I wouldn't make you do this outside the table for such a large number.)* We have fairly good statistical evidence students are more likely to be in a better mood on Mondays (but, we don't know why).

C. Estimate the proportion of USC students that feel better on Mondays than on Fridays, and give a 95% confidence interval for this proportion.

Wilson's estimator is:

$$\tilde{p} = 28/44 = 0.6363636,$$
$$\mathsf{SE}_{\tilde{P}} = \sqrt{0.6363636(1 - 0.6363636)/44} = 0.07252037$$

so a 95% CI is $0.6363636 \pm 1.96 \times 0.07252037 = (0.4942237, 0.7785035)$, i.e. from 49.4% to 77.9%.
Note: since this overlaps 50%, barely, this contradicts part B, that found $P < 0.05$ against a null of 50%. This difference is because B computes the P-value exactly, using the Binomial distribution, but the CI here uses the normal approximation. On a real test, I would (hopefully!) avoid contradictions like this.

D. Suppose we could not look up the P-value for the sign test on the table, but do have 40 fair coins. Briefly describe how we could use those coins to estimate the P-value in part B. *(this should take less than three sentences)*

Flip all 40 coins and write down the difference in number of heads and number of tails. Repeat this many times. The P value is estimated by the fraction of the times that we see at least 12 more heads than tails.

Suppose we have incomes (in dollars per year) from ten randomly sampled wife-husband pairs of LA residents:

wife	14000	17000	15000	35000	84000	0	19000	48000	23000	11000
husband	24000	17000	35000	0	98000	15000	12000	20000	196000	21000
difference	-10000	0	-20000	35000	-6000	-15000	7000	28000	-173000	-10000
signed rank	-3.5	(omit)	-6	+8	-1	-5	+2	+7	-9	-3.5

We would like to know if, in married heterosexual couples in LA, whether one gender tends to earn more money than the other.

1. Give two reasons why the Wilcoxon Signed-Rank test is a good choice for these data.

   We have paired samples; and the Wilcoxon test is better than a paired-sample t-test because the distribution of the values is very skewed (non-normal).

2. State the null and alternative hypotheses appropriate for carrying out a Wilcoxon Signed-Rank test.

   H_0 : The distribution of differences in income between husband and wife among LA couples is symmetric

   H_A : Either husbands, or wives, amongst LA couples, tend to make have higher incomes than the other.

3. Carry out the Wilcoxon Signed-Rank test.

   From the signed ranks above, $W_+ = 8+2+7 = 17$ and $W_- = 3.5+6+1+5+9+3.5 = 28$, so $W_s = 28$, and $n_D = 9$. (As a quick check, $W_+ + W_- = 45 = 9 \times (9-1)/2$.) The table gives $P > 0.2$.

4. What can we conclude about whether LA wives or husbands tend to have larger incomes?

   We have no strong statistical evidence that either wives or husbands tend to have higher incomes in LA, but our sample size was quite small, so there could still be a large effect that we did not detect.

Write the name of an appropriate statistical procedure – for instance "Wilcoxon–Mann–Whitney test" or "chi-squared for goodness of fit" or "t-test for difference in means".

A. We have measured the average flying speeds in five groups of 20 flies; each group was given a different number of milligrams of caffeine in their food. How much does the average flying speed increase, per milligram of caffeine?

t-test for no linear correlation/$r = 0$/slope $b_1 = 0$

B. We have measured heart rates in patients given combinations of dosages of two different drugs (none, low, and high doses for each, for a total of nine combinations). Is there an interaction between the drugs?

F-test for two-way ANOVA, for interactions

C. We have surveyed 100 randomly chosen individuals in Los Angeles and in Pasadena, and determined if they had a hospital visit in the last year. Does this differ between cities?

chi-squared test for a 2×2 contingency table

To study whether summer daylength causes human hair to absorb more light, we measured hair absorbance (larger numbers mean more absorbance) in random samples from four cities at different latitudes in Europe, summarized here:

	Barcelona	London	Paris	Stockholm
mean	4.41	5.04	4.73	5.52
SD	0.84	0.71	0.90	0.87
n	12	10	20	9

A. State the null hypothesis in words, and then in symbols.

H_0 : mean hair absorbance is the same in all cities

$\mu_{\text{Barcelona}} = \mu_{\text{London}} = \mu_{\text{Paris}} = \mu_{\text{Stockholm}}$

B. Construct the ANOVA table and test the null hypothesis.

Note: I will not make you calculate the SS(between), as we do below, on the test.

$$\begin{aligned}
\bar{\bar{y}} &= \frac{12 \times 4.41 + 10 \times 5.04 + 20 \times 4.73 + 9 \times 5.52}{51} = 4.854902 \\
\text{SS(within)} &= 11 \times 0.84^2 + 9 \times 0.71^2 + 19 \times 0.90^2 + 8 \times 0.87^2 = 27.80723 \\
\text{df(within)} &= 47 \\
\text{MS(within)} &= 0.5916432 \\
\text{SS(between)} &= 12 \times (4.41 - 4.85)^2 + 10 \times (5.04 - 4.85)^2 + 20 \times (4.73 - 4.85)^2 + 9 \times (5.52 - 4.85)^2 \\
&= 0.6901548 \\
\text{df(between)} &= 3 \\
\text{MS(between)} &= 0.2300516
\end{aligned}$$

$F_s = 0.2300516/0.5916432 = 0.388835$, so $P > .2$. Fail to reject H_0.

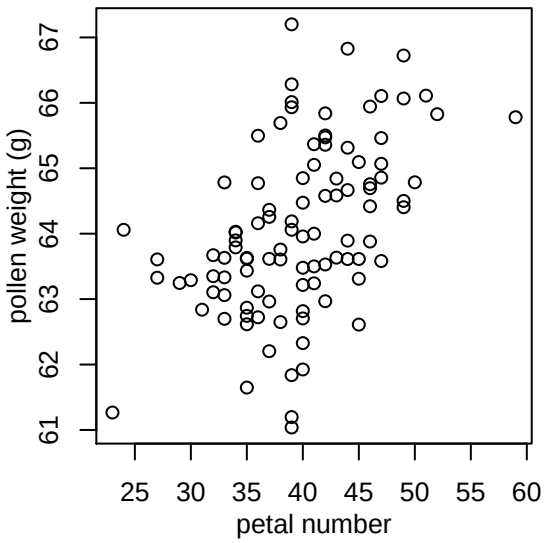
C. Estimate the within-group variability with the pooled standard deviation.

$$s_{\text{pooled}} = \sqrt{\frac{11 \times 0.84^2 + 9 \times 0.71^2 + 19 \times .9^2 + 8 \times .87^2}{47}} = 0.8473199,$$

D. Assess the strength of the evidence for and against the alternative hypothesis (i.e. what can we conclude).

We didn't find any evidence of a relationship between latitude and hair absorbance; however, there are confounding factors (genetics). Since the effect size we see is comparable to the within-group variability, we'd need much larger sample sizes, in any case.

It is hypothesized that flowers with more petals also produce more pollen. To test this, we have measured petal number and weight of pollen of 100 sunflowers, each from different plants, shown at right. The mean number of petals is 39.62, with an SD of 6.18; the mean weight of pollen is 64.05, with an SD of 1.29; and the correlation between them is $r = 0.49$.



A. Write the equation for the least-squares regression line, and draw this line on the plot.

Slope is $b_1 = .49 \times 1.29/6.18 = 0.1022816$; intercept is $b_0 = 64.05 - 0.1022816 \times 39.62 = 59.9976$.

$$\text{(pollen weight)} = 0.10 \times \text{(petal number)} + 60.0.$$

B. Formulate null and alternative hypotheses appropriate for a test of a linear relationship in words, and then in symbols.

H_0 : there is not a linear relationship between average petal number and average pollen weight

$b_1 = 0$

H_A : there is a linear relationship between average petal number and average pollen weight

$b_1 \neq 0$

C. Test the null hypothesis, and briefly state the conclusions.

$$t_s = .49\sqrt{\frac{98}{1 - .49^2}} = 5.564561$$

$$df = 98$$

and $P < .001$; we have strong evidence of a linear relationship between mean petal number and mean pollen weight.

D. If we found out that our field assistants had cut corners by collecting many of the sunflowers from the same plant, would this call our conclusions into question? Explain, briefly.

Yes, because the samples are not independent (flowers from the same plant could be correlated), so we might be overly confident in our results.