7. TWO-SAMPLE INFERENCE

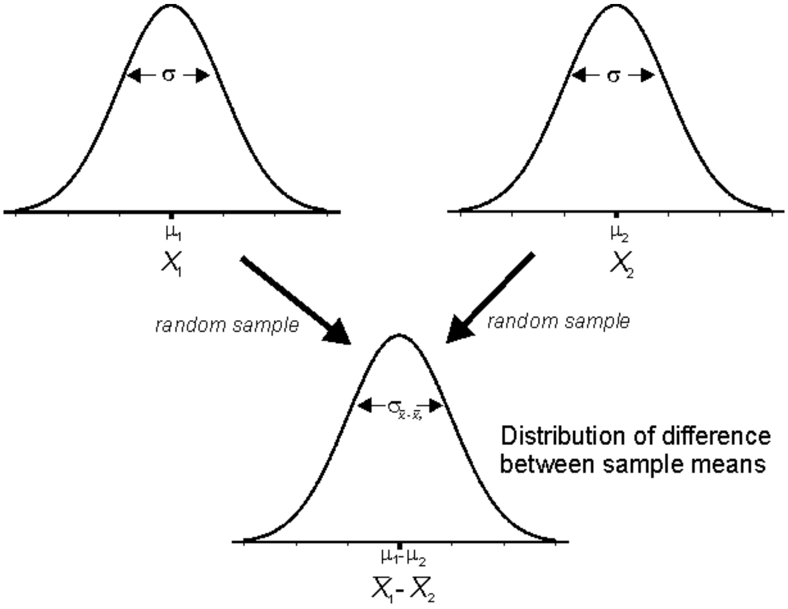
Two-sample vs. Paired-sample Designs

In this lab we consider the problem of estimating and testing differences between two means. For example, we may be interested in comparing the effects of two different medications on patient mean blood pressure. Or, we may wish to compare the effects of different fertilizers on mean plant growth. There are two completely different ways of carrying out such comparisons of means. The first approach is to randomly assign independent observations (e.g., patients, field plots) to different treatments. In this case we have **two samples** of individuals**, each from separate populations**: one sample of individuals given drug #1 and a second sample of individuals given drug #2 (or, one sample of field plots treated with fertilizer #1 and another sample treated with fertilizer #2). This is the **two-sample** design the subject of the present lab exercise. Our goal is to compare the two population means (μ_1 and μ_2) using two random samples of patients (or, field plots).

The second approach is to apply both treatments to each independent observation in the random sample (treat each patient with both drugs in random order and separated by time; or, divide each field plot into equal halves, and apply one fertilizer to one side and the second fertilizer to the other half). This is the "split plot" or **paired-sample** design, which is subject of next week's exercise.

Distribution of Differences Between Sample Means

The foundation for analysis of means of two populations is the fact that if X has a normal distribution in each of two populations, with equal variance σ^2, then the difference between sample means, $\overline{X}_1 - \overline{X}_2$, also has a normal distribution.


σ
σ
μ_1
X_1
μ_2
X_2
random sample
random sample
$\sigma_{\bar{x}_1-\bar{x}_2}$
Distribution of difference between sample means
$\mu_1-\mu_2$
X_1-X_2

You will have only a single estimate of each mean, but keep in mind that if you were to go back and collect two more random samples, the value of $\overline{X}_1 - \overline{X}_2$ obtained the second time would be different from that obtained the first time. The mean of the distribution of possible values for $\overline{X}_1 - \overline{X}_2$ is $\mu_1 - \mu_2$, and its standard deviation is $\sigma_{\overline{X}_1 - \overline{X}_2}$.

In this case, the quantity

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_{\overline{X}_1 - \overline{X}_2}}$$

has a *t*-distribution with $n_1 + n_2 - 2$ degrees of freedom. This fact is the basis of the two-sample *t*-test for a difference between population means, and of the confidence interval for the difference between two means. The quantity $s_{\overline{X}_1 - \overline{X}_2}$ is computed from the pooled sample variance, s_p^2, where

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

If *X* is normal in both populations with unequal variances, then a modified version of the above equation yields the Welch's *t*-statistic, which has an approximate *t*-distribution. Consult your textbook for this calculation, and for the calculation of the appropriate degrees of freedom.

Comparing Two Population Variances

The assumption of equal variances can be tested using the two-sample *F* test (JMP IN computes the Bartlett's test, which is exactly equivalent to the *F* test). This test is very sensitive to the assumption that the variable has a normal distribution in both populations. More robust, if less powerful, methods also exist including the Levene test, which tests for differences between populations in the mean absolute deviations. JMP IN computes the Levene test along with two other tests, the O'Brien's and the Brown-Forsythe tests.

Non-parametric Alternative to the Two-sample *t* Test

If the populations are not normally distributed, and sample size is not large enough to appeal to the Central Limit Theorem, then an alternative approach is to use a nonparametric test. Nonparametric tests are based on the ranks of the data rather than the data themselves, and they assume only that *X* is a continuous variable. The nonparametric equivalent of the two-sample *t*-test is the Wilcoxon rank sum test (equivalent to the Mann-Whitney *U* test). Under optimal conditions the Wilcoxon rank sum test is about 95% as powerful as a 2 sample *t*-test, although it may be less powerful in specific settings.

Power Analysis

When researchers carry out an experiment to test the difference between two treatment means, how do they decide on the appropriate sample sizes to take? How confident are they about their abilities to detect a difference if one is present? You haven't had to worry about this problem because we have provided the data sets and asked you to analyse them using the most appropriate procedures. But many of these data are from published studies that were designed intelligently: researchers decided on an appropriate sample size based in part on the expected power of the test. Power is the probability of correctly rejecting the null hypothesis when it is false (power is $1-\beta$, where β is the probability of making a Type II error). The power of the two-sample *t* test depends on:

1. The sample size (n_1+n_2). Greater sample size increases power of a test.
2. The significance level (α). Power decreases with decreasing α. For example, reducing α from 0.05 to 0.01 to reduce the probability of making a Type I error but increases the probability of making a Type II error.
3. The within-population variation (σ). Higher variation reduces power.
4. The difference between means, $\mu_1-\mu_2$. The larger the difference between the population means, the greater the probability of rejecting Ho.

In this lab we will explore the relationship between the power of the two-sample *t* test and these quantities.

Using The Program

The data must be entered into the data table as two separate columns. One column is a category (nominal) variable indicating treatment group (this will be "***X***"). The second column contains the actual measurements (this will be "***Y***"). Use **Fit Y by X** to start the analysis, placing the appropriate variables in the ***X*** and ***Y*** boxes. This will generate a "oneway" plot in which the observations of *Y* are displayed for each category of *X*. Click the red "▼" next to the "Oneway" title bar columns to select the following actions:

→ **Means/ANOVA/T-test**: Carries out the two-sample *t*-test; calculates a 95% confidence interval for difference between means; presents the Analysis of Variance (ANOVA) table [we will cover this method in a later lab]; adds the means diamonds to the plot (the vertical span of each diamond represents the 95% confidence interval for the mean of each group).

→ **UnEqual Variances**: Carries out tests of the null hypothesis that variances of the two populations are equal. We will use the Bartlett test (which is equivalent to the familiar two sample *F* test) and the Levene test. Also carried out are the calculations for the Welch's approximate *t*-test of differences between means when variances are unequal.

→ **Nonparametric → Wilcoxon Test**: Carries out the nonparametric analogue of the two-sample *t*-test. Note that exact *P*-values are not provided even when sample sizes are small: JMP IN uses the normal approximation regardless of sample size, and therefore provides only an approximate *P*-value, especially at small sample sizes.

→ **Power**: Power analysis of the two-sample *t*-test. It is most useful to vary only one of the quantities at a time. For example, select a range of sample sizes and leave the other quantities to their predetermined values. Click the **Solve for Power** and then **Done** to start computing. At the bottom of the output window there is an option to view the power curve. The only quantity you won't recognize is "Delta", which is a scaled measure of the difference between population means (see the Help features if you are interested in details); the preset value is calculated from the observed difference between sample means.

Problems

1. Dr. Jamie Smith, a professor in the Zoology Department at UBC, has studied song sparrows (*Melospiza melodia*) on the small Gulf island of Mandarte over several years. Mandarte Island is a short distance from Sidney, B.C., near the Victoria ferry terminal. Each summer for four years he captured every young song sparrow born on the island in that year, measured it, and placed a set of color bands on its legs. These bands uniquely identified each individual song sparrow. Each following spring Dr. Smith carried out a census of birds on Mandarte to determine which young had survived their first winter, and which had disappeared (presumed dead). A difference between survivors and dead birds in a trait would represent evidence of natural selection, the main cause of evolution according to Darwin. The data for young female birds is located in the file **song.sparrows.jmp**. Each line of the file refers to a single female bird. The variables, in order, are:

   - *Survival* - Whether the bird survived or died over her first winter
   - *mass* - Body mass, in g
   - *wing* - Wing length, in mm
   - *tarsus* - Tarsus ("leg") length, in mm
   - *beakL* - Beak length, in mm
   - *beakD* - Beak depth (height), in mm
   - *beakW* - Beak width, in mm

   a) Examine the distributions for beak length of surviving and dead sparrows (in the Distribution pop-up window, put *beakL* in the ***Y*** box and *Survival* in the **By** box; in the results window that appears, select ▼**Distributions→Stack** to display the two histograms one on top of the other). Do you notice a difference in the distributions of dead and surviving birds?

   b) Evaluate the fit of these two data sets to the normal distribution.

   c) A difference in the means of surviving and dead birds in a trait would reflect natural selection favoring one extreme over the other ("directional" selection). On the basis of your evaluation in (b), choose and carry out a test for a difference in mean beak length (*beakL*) between surviving and dead birds.

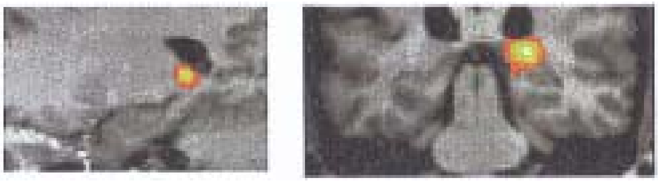
   d) What other assumption did your test in (c) require? Test this assumption with the beak length measurements. Was your assumption valid? What alternative methods are available if this assumption is not met?

   e) What is the 95% confidence interval for the difference between mean beak lengths of surviving and dead birds?

   f) Do surviving and dead birds differ in the means of any other traits? If so, do the larger individuals tend to survive better than the smaller birds in these traits as was the case for beak length?

g) Reduction in variance of survivors compared with dead birds, in the absence of a change in the mean, reflects a tendency for extreme individuals to do worse than individuals in the middle of the distribution (= “stabilizing” natural selection). Conversely, a higher variance among survivors than dead birds reflects a tendency for extreme individuals to do better than individuals in the middle of the distribution (= “disruptive” selection). Do any of the traits show evidence of stabilizing or disruptive selection?

h) Why is caution necessary when using the *F*-test (or, equivalently, the Bartlett test) for testing differences between populations in variance?



2. Maguire et al. (2000, Proc. Natl. Acad. Sci. USA 97: 4398–4403) used MRI to scan the brains of London taxi cab drivers, who are renowned for feats of spatial memory and navigation (individuals must undergo two years of extensive training and pass a stringent set of examinations known as “The Knowledge” before they can be licensed). MRI scans focussed on the hippocampus, a region of the brain associated with spatial memory (especially the posterior hippocampus). The data in the file **hippocampus.jmp** record the volume of gray matter (mm^3) in the right posterior hippocampus and the right anterior hippocampus of 15 drivers with different numbers of years of experience. Volume of the posterior hippocampus was measured using the VBM method, which provides a relative measure, whereas the anterior hippocampus was measured using a pixel-counting method that estimates absolute volume. All subjects were right-handed males between 32 and 62 years of age. These data were grabbed from Figure 3 in Maguire et al. (2000). The variables are:

a) Examine the difference in the volume of gray matter in the posterior hippocampus between the two experience groups of taxi drivers (< 15 years on the job vs. > 15 years on the job). Explain how you decided on the best method to use.

b) Repeat the above procedure on the anterior hippocampus measurements. Justify the methods you used.

c) Do the results of (a) and (b) imply that changes in the volume of gray matter in different regions of the hippocampus are influenced by experience as a London cab driver?

d) What does the following statement mean: “the two-sample *t*-test is more powerful than the Mann-Whitney *U*-test”?

e) Examine the influence of sample size on the power of the two-sample *t*-test, using the settings provided by the posterior hippocampus data. When $\alpha=0.05$ and σ and “Delta” match those of the posterior hippocampus data set, what sample size is needed to ensure that the power of the two-sample *t*-test is at least 0.5? What sample size is needed to ensure that the probability of rejecting Ho is at least 0.90?

f) What effect does reducing the significance level α to 0.01 instead of 0.05 have on the power of the two-sample *t*-test?

g) Do these data confirm that gray matter in different regions of the hippocampus change with number of years carrying out difficult feats of spatial memory?