Sample data

	Habitat 1	Habitat 2
Observation 1	7.6	5.9
Observation 2	0.4	3.8
Observation 3	1.1	6.5
Observation 4	3.2	18.3
Observation 5	6.5	18.2
Observation 6	4.1	16.1
Observation 7	4.7	7.6

Summary statistics

Statistic	Habitat 1	Habitat 2
ΣY_i	27.6	76.4
ΣY_i^2	150.52	1074.6
SS	41.70	240.75
γ	6	6
S^2	6.95	40.12
S	2.64	6.33
Mean ($\bar{Y}$)	3.94	10.91

Then calculate the F value as $F = S_1^2 / S_2^2 = 6.95 / 40.12 = 0.1732$

6) Compare the calculated value (0.1732) to the critical region. Given $\alpha = 0.05$ and a TWO TAILED alternative, and knowing that the degrees of freedom are $\gamma_1 = 6$ and $\gamma_2 = 6$, (note that both are equal), the critical limits are $P[0.1718 \leq F \leq 5.82] = 0.95$. Since our calculated F value is between these limit values we would fail to reject the null hypothesis, concluding that the data is consistent with the null hypothesis.

But it was close. Maybe there is a difference and we did not have enough power.

Some notes on F tests

NOTE that in this example the smaller value fell in the numerator. As a result, we were comparing the F value to the lower limit.

However, for two tailed tests, it makes no difference which falls in the numerator, and which in the denominator. As a result, we can ARBITRARILY decide to place the larger value in the numerator, and compare the F value to the upper limit.

The need to calculate the lower limit can be eliminated if we calculate $F = S^2_{\text{larger}} / S^2_{\text{smaller}}$.

However, don't forget that this arbitrary placing of the larger variance estimate in the numerator is done for TWO TAILED TESTS ONLY, and therefore we want to test against $F_{\alpha/2}$.

There are three common cases in F testing (actually two common and one not so common).

1) Frequently, particularly in ANOVA (to be covered later), we will test $H_0: \sigma_1^2 = \sigma_2^2$ against the alternative, $H_1: \sigma_1^2 > \sigma_2^2$. In this case we ALWAYS form the F value as $F = S_1^2 / S_2^2$.

We put the variance that is expected to be larger in the numerator for a one tailed test!

Don't forget that this is one tailed and all of α is placed in the upper tail. In the event that $F < 1$ we don't even need to look up a value in the table, it cannot be "significant".

2) Normal 2 tailed tests (used in 2 sample t-tests to be covered later) will test $H_0: \sigma_1^2 = \sigma_2^2$ against the alternative, $H_1: \sigma_1^2 \neq \sigma_2^2$. Here we can form the F value as $F = S^2_{\text{larger}} / S^2_{\text{smaller}}$.

Don't forget that this is a 2-tailed test and it is tested against the upper tail with only half of α (i.e. $\alpha/2$) in the upper tail.

When the larger value is placed in the numerator there is no way that we can get a calculated $F < 1$.

3) If both the upper and lower bounds are required (not common, found mostly on EXAMS in basic statistics) then we will be testing $H_0: \sigma_1^2 = \sigma_2^2$ against the alternative $H_1: \sigma_1^2 \neq \sigma_2^2$. We can form the F value any way we want, with either the larger or smaller variance in the numerator.

This is a 2 tailed test with $\alpha/2$ in each tail, and F can assume any positive value (0 to ∞)

Summary

The F distribution is ratio of two variances (i.e. two Chi square distributions) and is used test used to test two variances for equality. The null hypothesis is $H_0: \sigma_1^2 = \sigma_2^2$.

The distribution is an asymmetrical distribution with values ranging from 0 to ∞, and an expected value of 1 under the null hypothesis.

The F tables require two d.f. (numerator and denominator) and give only a very few critical values.

Many, perhaps most, F tests will be directional. For the tests the variance that is expected to be larger and hypothesized to be larger goes in the numerator whether it is actually larger or not. This value is tested against the upper tail with a probability equal to α.

For the non-directional alternative we may arbitrarily place the larger variance in the numerator and test against the upper tail, but don't forget to test against $\alpha/2$.

Probability Distribution interrelationships

The probability tables that we have been examining are interrelated. One of these interrelationships is actually pretty important!

If you examine the F table it turns out that in addition to F values the first column is equal to values of t^2, the last value in the first column corresponds to a Z^2 and the last row is a Chi square value divided by degrees of freedom.

d.f	1	2 ...	10	...	∞
1 2 . . . 10 . . .	t^2 (two tailed) (4.96)	F values		more F values	
∞	Z^2 (3.84)		γ_1=10 □ 1.83		1

The distributions and relationships we have discussed are:

1) $Z_i = \frac{Y_i - \mu}{\sigma}$ for observations and $Z = \frac{\bar{Y} - \mu_0}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}$ for testing hypothesis about means.

2) $\chi^2 = Z^2$ with 1 d.f.

$\chi^2 = \Sigma Z^2$ with n d.f.

$\chi^2 = SS/\sigma^2$ with n–1 d.f.

3) $t = \frac{\bar{Y} - \mu}{S_{\bar{Y}}} = \frac{\bar{Y} - \mu}{S / \sqrt{n}}$ with n–1 d.f.

4) $F = S_1^2 / S_2^2$ with n_1–1, n_2–1 d.f.

Interrelationships

1) χ^2/γ with γ d.f. = F with γ, ∞ d.f.

$$\frac{\chi^2}{\gamma} = \frac{\left(SS/\sigma^2\right)}{\gamma} = \left(\frac{\left(SS/\sigma^2\right)}{\gamma}\right)\left(\frac{1/\gamma}{1/\gamma}\right) = \left(\frac{SS}{\gamma\sigma^2}\right)\left(\frac{1/\gamma}{1/\gamma}\right) = \left(\frac{SS/\gamma}{\sigma^2}\right) = S^2/\sigma^2$$

which follows an F distribution with γ, ∞ d.f.

2) $F = \frac{\chi_1^2/\gamma_1}{\chi_2^2/\gamma_2}$ with γ_1, γ_2 d.f. if H_0 is true

given $\frac{\chi^2}{\gamma} = S^2/\sigma^2$ from part 1 above

then $\chi^2 = \gamma S^2 / \sigma^2$ and $\sigma^2\chi^2 = \gamma S^2$ and $S^2 = \sigma^2\chi^2 / \gamma$ with γ d.f.

therefore, $F = S_1^2 / S_2^2 = \sigma_1^2\chi_1^2/\gamma_1 \Big/ \sigma_2^2\chi_2^2/\gamma_2$ with $n_1 - 1$, $n_2 - 1 = \gamma_1, \gamma_2$ d.f.

if H_0 is true, then $\sigma_1^2 = \sigma_2^2$, then $F = \chi_1^2/\gamma_1 \Big/ \chi_2^2/\gamma_2$ with γ_1, γ_2 or $n_1 - 1$, $n_2 - 1$ d.f.

3) t with $\gamma = \infty$ follows a Z distribution, since as γ increases the sample variance (S^2) approaches the population variance (σ^2). That is, as the sample size approaches infinity the t distribution, $t = \frac{(Y_i - \bar{Y})}{S}$ approaches the Z distribution $Z = \frac{(Y_i - \mu)}{\sigma}$.

4) Z^2 = F with 1, ∞ d.f.

we saw that $Z^2 = \chi^2$ with 1 d.f.

we saw that χ^2/γ = F with γ, ∞ d.f.

then $F = \chi^2/\gamma = Z^2/\gamma = Z^2/1 = Z^2$.

5) t^2 with γ d.f. = F with 1, γ d.f.

This can be shown in several ways. First, we just saw that Z^2 = F with 1, ∞ d.f. This suggests that t^2 = F with 1, γ d.f. Another type of proof is given below.

$F = S_1^2 / S_2^2$ with γ_1, γ_2 d.f.

recall, $SS = \Sigma(Y_i - \bar{Y})^2$ with n d.f., and where $\hat{Y}_i$ is the mean of a subgroup of the data partitioned into two (or more) groups (the basis of Analysis of Variance).

$$SS = \Sigma(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \bar{Y} + \hat{Y}_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}\left((\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)\right)^2$$

d.f are n = (n – 1) + 1

Let $S_1^2 = n(\bar{Y} - \mu)^2$ with 1 d.f.

Let $S_2^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 / (n-1)$ with n–1 d.f.

both of which are unbiased estimates,

then $F = S_1^2 / S_2^2 = n(\bar{Y} - \mu)^2 / S_2^2$ with 1, n–1 d.f.

and $F = \frac{n(\bar{Y} - \mu)^2}{S_2^2} = \frac{(\bar{Y} - \mu)^2}{S_2^2/n} = \left(\frac{(\bar{Y} - \mu)}{S_2/\sqrt{n}}\right)^2 = t^2$ with n–1 d.f.

Summary

1) χ^2/γ with γ d.f. = F with γ, ∞ d.f.

2) $F = \frac{\chi_1^2}{\gamma_1} \Big/ \frac{\chi_2^2}{\gamma_2}$ with γ_1, γ_2 d.f. if Ho is true

3) t with $\gamma = \infty$ follows a Z distribution

4) Z^2 = F with 1, ∞ d.f.

5) t^2 with γ d.f. = F with 1, γ d.f.

Some Examples in the F tables (all $\alpha = 0.05$, two tails for Z and t values since sign is lost in squaring)

F with 1, 10 d.f. = 4.96 = t^2 with 10 d.f. = $(2.228)^2$ = 4.96

F with 1, ∞ d.f. = 3.84 = Z^2 = $(1.96)^2$ = 3.84

F with 10, ∞ d.f. = 1.83 = χ^2 / γ = 18.3 / 10 = 1.83 with 10 d.f.

Confidence intervals and margin of error

The confidence interval is an expression of what we believe to be a range of values that is likely to contain the true value of some parameter is called a confidence interval. The width of this interval above and below the parameter estimate is called the margin of error.

We can calculate confidence intervals for means (μ) and variances (σ^2).

Confidence intervals for t and Z distributions

t and Z distribution confidence intervals start with a t or Z probability statement.

$P(-t_{\alpha/2} \le t \le t_{\alpha/2}) = 1 - \alpha$ can also be written

$$P(-t_{\alpha/2} \le \frac{\bar{Y} - \mu}{S_{\bar{Y}}} \le t_{\alpha/2}) = 1 - \alpha$$

which is modified to express an interval about μ instead of t (or Z).

$$P(-t_{\alpha/2} S_{\bar{Y}} \le \bar{Y} - \mu \le t_{\alpha/2} S_{\bar{Y}}) = 1 - \alpha$$

$$P(-\bar{Y} + t_{\alpha/2} S_{\bar{Y}} \ge -\mu \ge -\bar{Y} - t_{\alpha/2} S_{\bar{Y}}) = 1 - \alpha$$

The final form is given below.

$$P(\bar{Y} - t_{\alpha/2} S_{\bar{Y}} \le \mu \le \bar{Y} + t_{\alpha/2} S_{\bar{Y}}) = 1 - \alpha$$

The expression for Z has an identical derivation.

$$P(\bar{Y} - Z_{\alpha/2} \sigma_{\bar{Y}} \le \mu \le \bar{Y} + Z_{\alpha/2} \sigma_{\bar{Y}}) = 1 - \alpha$$

A common short notation for the interval in the probability statement is given as $\overline{Y} \pm t_{\alpha/2} S_{\overline{Y}}$, but the probability statement is preferable as a final result. The value $t_{\alpha/2} S_{\overline{Y}}$ for intervals on means and $t_{\alpha/2} S$ for intervals on individual observations is half of the interval width from the lower limit to the upper limit and is called the margin of error.

Confidence intervals for variance

Variances follow a Chi square distribution. The confidence interval for variance is based on the Chi Square distribution.

$$P(\chi^2_{lower} \le \chi^2 \le \chi^2_{upper}) = 1 - \alpha \text{ or}$$

$$P(\chi^2_{lower} \le \frac{SS}{\sigma^2} \le \chi^2_{upper}) = 1 - \alpha$$

which is solved to isolate σ^2.

$$P\left(\frac{1}{\chi^2_{lower}} \ge \frac{\sigma^2}{SS} \ge \frac{1}{\chi^2_{upper}}\right) = 1 - \alpha$$

$$P\left(\frac{1}{\chi^2_{upper}} \le \frac{\sigma^2}{SS} \le \frac{1}{\chi^2_{lower}}\right) = 1 - \alpha$$

giving the expression,

$$P\left(\frac{SS}{\chi^2_{upper}} \le \sigma^2 \le \frac{SS}{\chi^2_{lower}}\right) = 1 - \alpha$$

Notice that the upper tabular Chi square value comes out in the lower bound and the lower Chi square in the upper bound.

$$P\left(\frac{SS}{\chi^2_{upper}} \le \sigma^2 \le \frac{SS}{\chi^2_{lower}}\right) = 1 - \alpha$$

Notes on confidence intervals

One sided intervals are possible, but uncommon.

Confidence intervals are one of the most common expressions in statistics, frequently occurring in publications.

Margins of error and confidence intervals are not always calculated in statistical software programs, but they can easily be done by hand.

From the previous SAS Example 2c

We receive a shipment of apples that are supposed to be "premium apples", with a diameter of at least 2.5 inches. We will take a sample of 12 apples, and place a confidence interval on the mean. The sample values for the 12 apples are;

2.9, 2.1, 2.4, 2.8, 3.1, 2.8, 2.7, 3.0, 2.4, 3.2, 2.3, 3.4

Do we want the Std dev or Std error?

SAS PROC UNIVARIATE Output

The UNIVARIATE Procedure
Variable: diam (Diameter of the apple)

Moments

N	12	Sum Weights	12
Mean	2.75833333	Sum Observations	33.1
Std Deviation	0.39418116	Variance	0.15537879
Skewness	-0.1184219	Kurtosis	-0.8352969
Uncorrected SS	93.01	Corrected SS	1.70916667
Coeff Variation	14.2905557	Std Error Mean	0.1137903

The standard deviation is the variation in individual apples. If we wanted the interval that contained 95% of the apples, we would use the standard deviation. However, we have estimated a mean and we want to place a confidence interval that expresses our knowledge of this estimate. Is our estimate of the mean good or poor? Is the confidence interval narrow or wide?

Note the confidence interval about the mean, and about individual observations.

For means: $P(\bar{Y} - t_{\alpha/2} S_{\bar{Y}} \le \mu \le \bar{Y} + t_{\alpha/2} S_{\bar{Y}}) = 1 - \alpha$

The margin of error for the mean is $t_{\alpha/2} S_{\bar{Y}}$

For individual observations: $P(\bar{Y} - t_{\alpha/2} S \le \mu \le \bar{Y} + t_{\alpha/2} S) = 1 - \alpha$

So we need the mean of the apples and the standard error.

Mean = 2.758333

Std Error Mean = 0.11379 (no adjustment needed)

We also need a t–value. With 12 apples and 11 d.f., our two tailed t–value is 2.201. So $\bar{Y} \pm t_{\alpha/2} S_{\bar{Y}}$ or 2.758 ± (2.201)(0.1138) = 2.758 ± (0.250) gives the interval. The margin of error is 0.250 and the best expression is as a confidence interval probability statement.

P(2.758 – 0.250 ≤ μ ≤ 2.758 + 0.250) = 1 – α

P(2.508 ≤ μ ≤ 3.008) = 0.95

The real value of μ may or may not be in this interval, but it is our best evaluation of where the true value of μ will be.

For individual observations the calculation uses the standard deviation instead of the standard error, $P(\bar{Y} - t_{\alpha/2} S \le \mu \le \bar{Y} + t_{\alpha/2} S) = 1 - \alpha$. For the apples the standard deviation was 0.3942 and all other values remain the same.

$\bar{Y} \pm t_{\alpha/2} S$ or 2.758 ± (2.201)(0.3942) = 2.758 ± (0.8676) gives the interval. The best expression is as a confidence interval probability statement.

P(2.758 – 0.8676 ≤ μ ≤ 2.758 + 0.8676) = 1 – α

P(1.8904 ≤ μ ≤ 3.6256) = 0.95

Example 2 - Variance CI

Place a confidence interval on the variance estimate for the apple example. The variance estimate from the SAS output is $S^2 = 0.155379$ and the corrected sum of squares is 1.709. The Chi square values for 11 d.f. are 3.816 (lower) and 21.92 (upper).

Recall, $P\left(\frac{SS}{\chi^2_{upper}} \le \sigma^2 \le \frac{SS}{\chi^2_{lower}}\right) = 1-\alpha$

Then $P\left(\frac{1.709}{21.92} \le \sigma^2 \le \frac{1.709}{3.816}\right) = 0.95$ and $P\left(0.078 \le \sigma^2 \le 0.448\right) = 0.95$,
for a variance of $S^2 = 0.155379$ and a corrected SS of 1.709.

A note on hypothesis testing

Hypothesis tests can be done by calculating a confidence interval for the appropriate value of α and checking to see if the hypothesized value is contained in the interval. This approach is used in some SAS program output such as Analysis of Variance.

Summary

Confidence intervals for μ (t or Z distribution) and σ^2 (Chi square).

For μ (using either the t or Z distribution).

$$P(\bar{Y} - t_{\alpha/2} S_{\bar{Y}} \le \mu \le \bar{Y} + t_{\alpha/2} S_{\bar{Y}}) = 1-\alpha$$

Where the margin of error for the mean is $t_{\alpha/2} S_{\bar{Y}}$ or $Z_{\alpha/2} S_{\bar{Y}}$ for Z distribution applications.

For σ^2 (Chi square distribution).

$$P\left(\frac{SS}{\chi^2_{upper}} \le \sigma^2 \le \frac{SS}{\chi^2_{lower}}\right) = 1-\alpha$$

These are common and IMPORTANT calculations.

The margin of error is the amount added and subtracted from the mean to get the confidence interval. It is equal to $t_{\alpha/2} S_{\bar{Y}}$.

They are not always calculated by statistical software.

Checking to see if a value falls in the interval is equivalent to perform statistical tests of hypothesis against that value.

Linear combinations

Generic Example: We want to create a score we can use to evaluate students applying to LSU as freshmen.

$$Score_i = a\left(VerbalSAT_i\right) + b\left(MathSAT_i\right) + c\left(GPA_i\right)$$

where; a, b and c are constants and

VerbalSAT, MathSAT and GPA are variables that vary among students.

We need to choose values of a, b and c in order to calculate the score. The average student has values of VerbalSAT=500, MathSAT=500 and GPA=2.

If we choose a=1/3 and b=1/3 and c=1/3 then we have an average of the 3 variables, (a+b+c)/3. The score for the "average" student would be 334.1633. If we choose a=1 and b=1 and c=1 we have a simple sum of the variables, (aV+bM+cG). For the average student this would be 1002.5. Both of these choices produce a score that could be used to compare among students. However, neither of these two choices produces a score that resembles the more familiar scores for standardized tests or grade point averages.

But VerbalSAT and MathSAT are values in the hundreds (range: 200 to 800) and the GPA is single digits (range: 0 to 4). If we wanted to scale our scores to more closely resemble the mean of these scores we could modify the coefficients for means (a=1/3 and b=1/3 and c=1/3) by either dividing the standardized tests by an additional 200 points so their maximum would be 4 (a=1/600 and b=1/600 and c=1/3) producing a mean for the average student of 2.5 or we could scale the GPA by multiplying by 200 to produce a maximum of 800 (a=1/3 and b=1/3 and c=200/3) yielding a score of 500 for the average student.

Any of these choices produce a reasonable and acceptable linear combination. Which one is used depends on the objectives and preferences of the user. We will look at several specific applications.

Mean and Variance of a linear combination

So what is the mean value of our linear combination, and can we put a variance on it (to get a confidence interval)?

$$Score_i = a_1\left(VerbalSAT_i\right) + a_2\left(MathSAT_i\right) + a_3\left(GPA_i\right) = \sum_{i=1}^{3} a_i Y_i$$

Linear combination: $\sum a_i Y_i$

Expected value: $\sum a_i \mu_{Y_i}$

Estimate of mean: $\sum a_i \overline{Y}_i$

Variance of the linear combination.

The variance of a linear combination is the sum of the individual variances (with squared coefficients) plus twice the covariance of the variables (with both coefficients).

Estimate of the variance:

$$\sum a_i^2 S_i^2 + \sum\sum 2a_i a_j S_{ij} \text{ for i} \neq \text{j}$$

For example, with the Score we calculated previously, the variance might be

$$VAR(Score) = a^2 VAR(Verbal) + b^2 VAR(Math) + c^2 VAR(GPA) + 2abCov(Verbal, Math) + 2acCov(Verbal, GPA) + 2bcCov(Math, GPA)$$

HOWEVER, if the variables are independent the covariance can be assumed to be zero. The linear combination reduces to the sum of the variances of the individual variables (with squared coefficients).

$$VAR(Score) = a^2 VAR(Verbal) + b^2 VAR(Math) + c^2 VAR(GPA)$$

Utility of linear combinations

As the course progresses we will see applications of linear combinations to almost everything.

Two sample t-test: $H_0: \mu_1 - \mu_2 = \delta$, estimated by $\bar{Y}_1 - \bar{Y}_2 = \delta$. This is a linear combination! The most common case is a test of the linear combination $1\bar{Y}_1 + (-1)\bar{Y}_2 = \delta$, but any other combination is possible (e.g. $0.8\bar{Y}_1 - 1\bar{Y}_2 = 0$) and we only need calculate a variance and standard error for the given situation.

$Var\left(0.8\bar{Y}_1 - 1\bar{Y}_2\right) = (0.8)^2 S^2_{\bar{Y}_1} + (-1)^2 S^2_{\bar{Y}_2} = 0.64 S^2_{\bar{Y}_1} + S^2_{\bar{Y}_2}$.

Regression: The model, $Y_i = b_0 + b_1 X_i + e_i$ is a linear combination.

Analysis of variance: We will look at contrasts to test for differences between means similar to a two sample t-test (but usually with more than two means). For example, testing the hypothesis that some mean is equal to the average of two other means would be done as $H_0: \mu_1 - \left(\frac{\mu_2 + \mu_3}{2}\right) = 1\mu_1 + \left(\frac{-1}{2}\right)\mu_2 + \left(\frac{-1}{2}\right)\mu_3 = 0$

An application: Stratified Random Sampling

Suppose we want to estimate the number of ducks in an area on the Louisiana coast. The area of interest is 300 acres. We fly 9 transects, counting ducks for 1/10 mile on either side of the plane, and from each transect we estimate the number of ducks per acre. We can then calculate an estimate and a confidence interval for that estimate.

Raw data

Sample Number	Ducks counted
1	8
2	19
3	30
4	23
5	56
6	89
7	2024
8	1732
9	1122

Summary statistics

Statistic	Value
n =	9
sum =	5103
mean =	567
var =	684349.25
std dev =	827.25
std error =	275.75
t-value =	2.306
acres =	300

Estimate and confidence interval for the (300 acres)

Given this value as the estimate of the true mean number of ducks (μ), we can state our results as a probability statement. The usual calculation is

"*Some parameter estimate* ± $t_{\alpha/2}$ * *standard error*"

The confidence interval for ducks per acre is $\bar{Y} \pm t_{\alpha/2} S_{\bar{Y}}$ or $567 \pm 2.306(275.75)$ and

567 ± 635.88. However, we want ducks per 300 acres. Recall from our discussion of transformations that when the mean is multiplied by 300 to get total ducks on the 300 acres, the variance would be multiplied by 300 squared and the standard deviation multiplied by 300.

So the calculation is $(300)567 \pm 2.306(300)(275.75)$, so the margin of error is

$2.306(300)(275.75) = 190765$ and the interval is 170100 ± 190765

The probability statement is

$$P(\bar{Y} - t_{\alpha/2} S_{\bar{Y}} \le \mu_{ducks} \le \bar{Y} + t_{\alpha/2} S_{\bar{Y}}) = 1 - \alpha$$

$$P(-20665 \le \mu \le 360865) = 0.95$$

This calculation is for the number of ducks is the value of interest and the 300 acres the area of interest.

Stratification

Now let's suppose we noticed that the duck species we were studying and counting were primarily a fresh water species, and occurred only infrequently in the brackish and saline zones of the 300 acres of interest. We could modify the study and examine the numbers in the 3 zones separately. The advantage is that perhaps we can get 3 separate estimates of homogeneous habitat with a smaller variance than one estimate of the heterogeneous whole 300 acres. We will allow that each zone has been determined to encompass 100 acres to simplify our analysis.

The 3 habitat types would be called "strata", and we could estimate the number for each stratum separately and, we assume, independently so we need not consider covariance.

The samples were done such that there are 3 samples in each stratum.

Obs	Saline (CH_2O)	Brackish (BH_2O)	Fresh (FH_2O)
1	8	23	2024
2	19	56	1732
3	30	89	1122
n =	3	3	3
sum =	57	168	4878
mean =	19	56	1626
var =	121	1089	211828
std dev =	11	33	460.248
std error =	6.35	19.05	265.72

$$\overline{Y}_{Total\ Ducks} = 100\overline{Y}_{FH_2O} + 100\overline{Y}_{BH_2O} + \overline{Y}_{CH_2O}$$

$$Total\ Ducks = 100(1626) + 100(56) + 100(19) = 1900 + 5600 + 162600 = 170100$$

The estimated total numbers were the same since each the area of each zone was the same.

Now we need a variance of the mean in order to calculate a confidence interval.

$$S^2_{\overline{Y}} = a^2 S^2_{\overline{Y}} + b^2 S^2_{\overline{Y}} + c^2 S^2_{\overline{Y}}$$

$$S^2_{\overline{Y}_{TotalDucks}} = a^2 S^2_{\overline{Y}_{FH_2O}} + b^2 S^2_{\overline{Y}_{BH_2O}} + c^2 S^2_{\overline{Y}_{CH_2O}}$$

$$S^2_{\overline{Y}_{TotalDucks}} = (100)^2\, 211828 + (100)^2\, 1089 + (100)^2\, 121 = 2130380000$$

$$S_{\overline{Y}_{TotalDucks}} = \sqrt{2130380000} = 15385.347$$

Linear combination of independent means.

t-value	2.306
acres	300
estimated total	170100
variance	2,130,380,000
std error	15385.347
margin of error	35478.70
CL-lower	134621.30
CL-upper	205578.70

Old and new estimates.

P(–20665 ≤ Total Ducks ≤ 360865) = 0.95

P(134621 ≤ Total Ducks ≤ 205579) = 0.95

Why does stratification give a smaller interval width? Because we replaced one sample with a very large variance (684349) with 3 samples, each with a smaller variance (121, 1089 and 211828). This reduced the margin of error from 190,765 to 35,478.

The heterogeneous variances should not have been pooled in the first place.

You will recall that we mentioned that one way of increasing power is to reduce the variance. We did not dwell on this previously because the only mechanism I could suggest at the time for reducing variance was "improving measurement error". Now we have another method of reducing variance, stratification. This involves sampling smaller homogeneous units instead of one large heterogeneous unit.

Is the fact that the 3 variances were not similar a problem? No, nowhere in working with linear combinations did we state that the variances had to be similar. Later we will find that this can be advantageous, but it is not necessary for this type of analysis. It is advantageous in analyses where we may want to pool variances to a single, improved estimate.

Summary

The use of "Linear Combinations" is a rather generic technique with many applications throughout statistics. We will see them again in the two sample t-test, regression, and ANOVA. Sampling is another example of the application.

It is important to determine if the variables in the linear combination are independent or not. If they are, the covariance values can be considered to be zero.

The linear combination and its variance is calculated as

Linear combination = $\sum_{i=1}^{k} a_i Y_i$

Variance = $\sum_i a_i^2 S_i^2 + 2\sum_i \sum_j a_i a_j S_{ij}$ for i ≠ j

Where, if we can assume the variables are independent the covariances can be assumed to be zero. This will be very important later. We will assume independence in t-test and ANOVA, and parts of regression, but not all of regression!

A final note on linear combinations

We have been assuming "independence" for a while. We sample at random to obtain independence, and to get a good representative sample. But do "linear combinations" have anything to do with the simpler calculations we talked about earlier when we assumed independence, say a test of the mean against an hypothesized value?

I'm glad you asked. As a matter of fact the mean is calculated as $\bar{Y} = \left(Y_1 + Y_2 + Y_3 + ... + Y_n\right) / n$

and that is a linear combination. Fortunately for us we do not have to consider the covariances of individual observations because they are sampled independently!