Session 6: Further single and two and more sample statistics

Confidence intervals

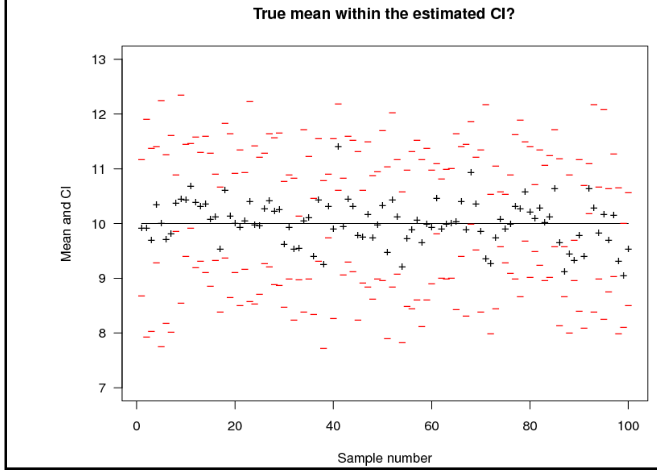
Confidence intervals

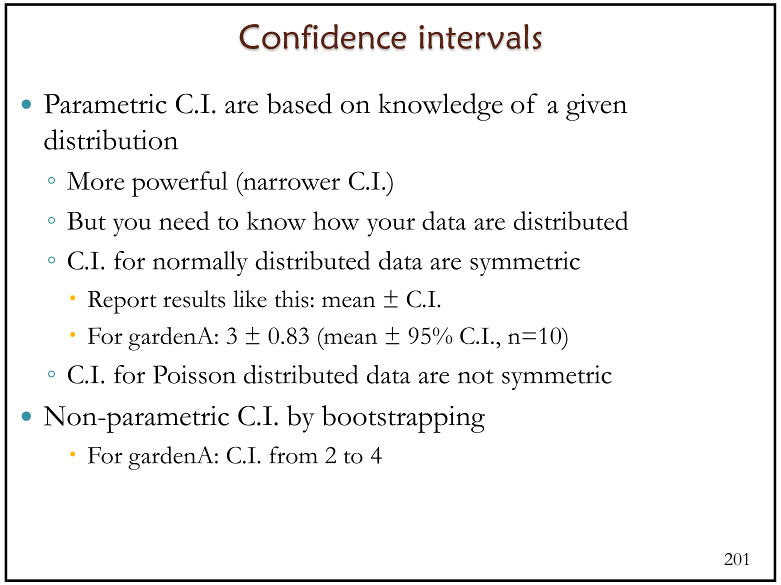
- Let's do a thought experiment on confidence intervals, where we know the true mean (population mean)
  - We repeat an experiment 100 times and each time calculate the mean and confidence interval from this sample
    - Of course the mean and C.I. will differ from sample to sample (estimates!)
  - We find that the true mean is within the range spanned by the confidence interval most of the time but not always
    - For a 95% C.I. you can expect that the true mean is within 95 of the 100 times you repeated the experiment
- The 95% confidence interval is that range that includes the true mean in 95% of the cases (of doing the experiment)
  - For normally distributed data, the confidence interval is the standard error of the mean scaled by the *critical-t* (with particular α and d.f.)
  - C.I. = t · s.e.m.

$$CI_{95\%} = t_{(\alpha=0.025, d.f.=?)} \sqrt{\frac{s^2}{n}}$$

Confidence intervals

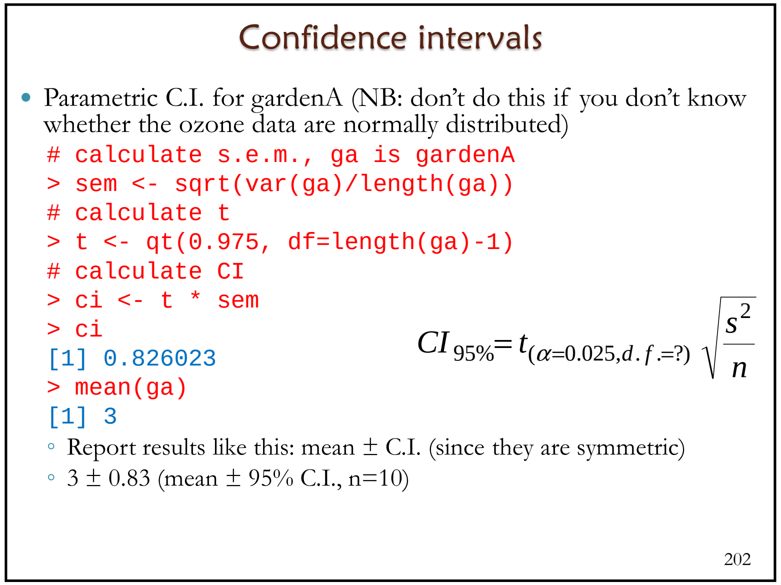
- Results of this thought experiment in R





Confidence intervals

- Parametric C.I. are based on knowledge of a given distribution
  - More powerful (narrower C.I.)
  - But you need to know how your data are distributed
  - C.I. for normally distributed data are symmetric
    - Report results like this: mean ± C.I.
    - For gardenA: 3 ± 0.83 (mean ± 95% C.I., n=10)
  - C.I. for Poisson distributed data are not symmetric
- Non-parametric C.I. by bootstrapping
    - For gardenA: C.I. from 2 to 4



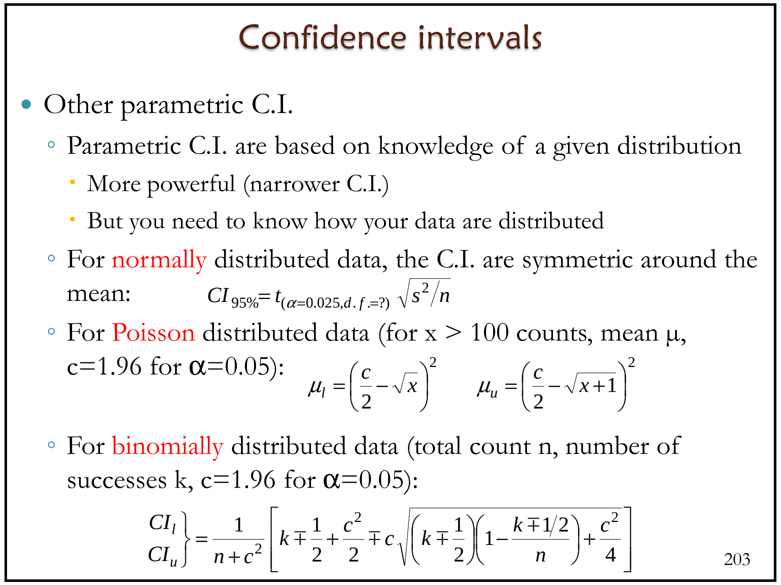
Confidence intervals

- Parametric C.I. for gardenA (NB: don't do this if you don't know whether the ozone data are normally distributed)

```
# calculate s.e.m., ga is gardenA
> sem <- sqrt(var(ga)/length(ga))
# calculate t
> t <- qt(0.975, df=length(ga)-1)
# calculate CI
> ci <- t * sem
> ci
[1] 0.826023
> mean(ga)
[1] 3
```

$$CI_{95\%} = t_{(\alpha=0.025,d.f.=?)}\sqrt{\frac{s^2}{n}}$$

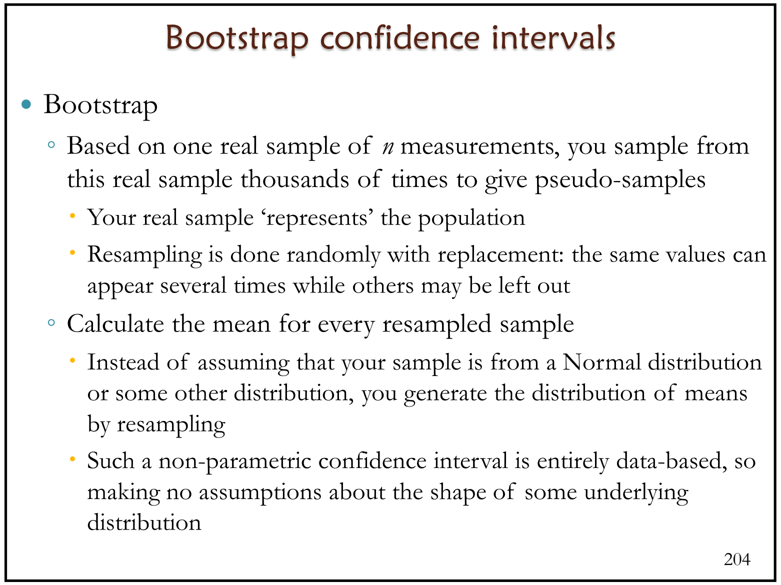
  - Report results like this: mean ± C.I. (since they are symmetric)
  - 3 ± 0.83 (mean ± 95% C.I., n=10)



Confidence intervals

- Other parametric C.I.
  - Parametric C.I. are based on knowledge of a given distribution
    - More powerful (narrower C.I.)
    - But you need to know how your data are distributed
  - For normally distributed data, the C.I. are symmetric around the mean: $CI_{95\%} = t_{(\alpha=0.025,d.f.=?)}\sqrt{s^2/n}$
  - For Poisson distributed data (for x > 100 counts, mean μ, c=1.96 for α=0.05): $\mu_l = \left(\frac{c}{2} - \sqrt{x}\right)^2$ $\mu_u = \left(\frac{c}{2} - \sqrt{x+1}\right)^2$
  - For binomially distributed data (total count n, number of successes k, c=1.96 for α=0.05):

$$\left.\begin{matrix} CI_l \\ CI_u \end{matrix}\right\} = \frac{1}{n+c^2}\left[k \mp \frac{1}{2} + \frac{c^2}{2} \mp c\sqrt{\left(k \mp \frac{1}{2}\right)\left(1 - \frac{k \mp 1/2}{n}\right) + \frac{c^2}{4}}\right]$$

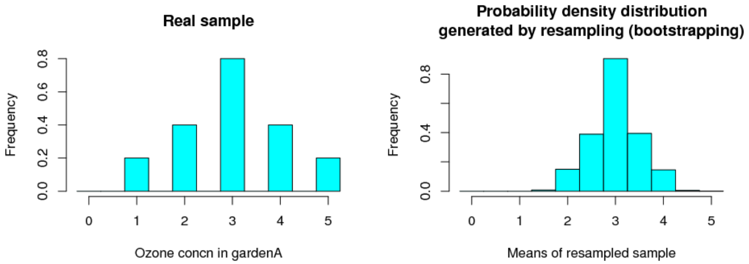


Bootstrap confidence intervals

- Bootstrap
  - Based on one real sample of *n* measurements, you sample from this real sample thousands of times to give pseudo-samples
    - Your real sample 'represents' the population
    - Resampling is done randomly with replacement: the same values can appear several times while others may be left out
  - Calculate the mean for every resampled sample
    - Instead of assuming that your sample is from a Normal distribution or some other distribution, you generate the distribution of means by resampling
    - Such a non-parametric confidence interval is entirely data-based, so making no assumptions about the shape of some underlying distribution

Bootstrap confidence intervals

- Bootstrapping of some data



Bootstrap confidence intervals

- For-loop (simple examples to show what a for-loop does)
  - For repetitive tasks, like resampling 10,000 times

```
for (i in 1:10) {
  print(i)
}

v <- runif(10)
for (i in 1:10) {
  print(i)
  print(v[i])
}
```

Bootstrap confidence intervals

- Non-parametric C.I. by bootstrapping of gardenA data

```
# m will contain 10,000 means of resampled data
m <- numeric(10000)
for (i in 1:10000) {
  m[i] <- mean(sample(ga, 5, replace=TRUE))
}
mean(m)
[1] 2.99562
quantile(m, c(0.025,0.975))
 2.5% 97.5%
    2     4
```

  - Compare with parametric C.I. results:
  - 3 ± 0.83 (mean ± 95% C.I., n=10)

Exercise

- Calculate the C.I. for gardenB
  - Download the text file gardens.txt from WebCT and read into R
  - Make a box-and-whisker plot of gardenB data
  - Calculate C.I. assuming data are normally distributed
  - Calculate C.I. using bootstrapping and compare with above
  - Plot a histogram of all means from bootstrapping (means of all samples generated by random picking from real sample)

More tests

- F-test for significance of differences in variance
- Tests for how data are distributed, e.g. normality
  - Kolmogorov-Smirnov (KS) test
- Non-parametric versus parametric tests
- Measurement scales
- Non-parametric alternatives to the t-test
  - Wilcoxon's rank sum test (signed rank test)

Fisher's F-test

- F-test: are the variances between two samples significantly different?
  - Assumption: normally distributed data
- Example: gardenB and gardenC have the same mean but different variance

```
> mean(gb)
[1] 5
> mean(gc)
[1] 5
> var(gb)
[1] 1.333333
> var(gc)
[1] 14.22222
```

Fisher's F-test

- F-statistic: the ratio of the two variances (let's divide larger variance by smaller variance so F > 1)

```
> f <- var(gc)/var(gb)
[1] 10.66667
```

- The critical value of this variance ratio can be obtained from the quantile function for the F-distribution qf()
  - For a two-tailed test at $\alpha=0.05$ we use q=1-0.025
  - The arguments for qt() are this quantile 0.975, d.f. for numerator, d.f. for denominator (in that order)

```
> qf(0.975, length(gc)-1, length(gb)-1)
[1] 4.025994
```

- The test statistic (F=10.7) is > the critical value, so the sample variances are significantly different

Fisher's F-test

- Now calculate the p-value from the probability of observing an F-statistic like this or more extreme
  - The summed probabilities of all F-statistic values up to our F value of 10.7 can be taken from the cumulative F-distribution pf() where the arguments are F-statistic, d.f. for numerator, d.f. for denominator (in that order)

```
> pf(f, length(gc)-1, length(gb)-1)
[1] 0.999188
```

Fisher's F-test

- The probability of the F-statistic being 10.7 or more extreme is the complement

```
> 1 - pf(f, length(gc)-1, length(gb)-1)
[1] 0.0008120995
```

- This is the probability of F being larger than our 10.7, but we should use a two-tailed F-test so we multiply by 2

```
> cf <- 2*(1 - pf(f, length(gc)-1, length(gb)-1))
[1] 0.001624199
```

- Now this is our p-value, it is well below 0.05 so we reject the null hypothesis that the variances are equal
- The p-value of 0.0016 is the probability of observing a variance ratio (F-ratio or F-statistic) as extreme or more extreme than 10.7 if H_0 were correct

Fisher's F-test

- Now let's use the ready-made R function var.test() (it doesn't matter in which sequence you enter the samples)

```
> var.test(gc,gb)
        F test to compare two variances
data:  gc and gb
F = 10.6667, num df = 9, denom df = 9, p-value =
  0.001624
alternative hypothesis: true ratio of variances is not
  equal to 1
95 percent confidence interval:
  2.649449 42.943938
sample estimates:
ratio of variances
          10.66667
```

Exercise

- Are the variances between gardenA and gardenB data significantly different?
  - Download gardens.txt file from WebCT and read in as described previously
  - Do an F-test by hand and by the ready-made function
  - Compare with results from gardenB and gardenC data

Tests for Normality: Kolmogorov-Smirnov (KS) test

- Not just for testing for Normality! Can test two related types of questions:
  - Are the two sample distributions the same?
  - Does a particular sample distribution come from a particular theoretical distribution, e.g. from a Normal distribution, or from a log-normal distribution, etc.?
  - The differences between sample distributions or sample and theoretical distribution could be due to differences in
    - mean (see t-test)
    - variance (see F-test)
    - skew
    - etc.

Kolmogorov-Smirnov (KS) test

- The Kolmogorov-Smirnov test works by comparing the cumulative probability functions of the two distributions under consideration
  - Example data again from Crawley's website, sizes of insect wings in two different locations (wings.txt)

```
> w <- read.table("wings.txt",header=TRUE)
> attach(w)
> names(w)
[1] "size"     "location"
> table(location)
location
 A  B
50 70
```

Kolmogorov-Smirnov (KS) test

- Separate the wing sizes into two samples by location

```
> A <- size[location == "A"]
> B <- size[location == "B"]
```

- Let's compare the two sample distributions with ks.test()

```
> ks.test(A,B)
        Two-sample Kolmogorov-Smirnov test
data:  A and B
D = 0.2629, p-value = 0.02911
alternative hypothesis: two-sided
```

  - Given that p-value < 0.05 the two distributions are significantly different, but in what respect?

Kolmogorov-Smirnov (KS) test

- Maybe the means are different? Let's do a t-test (Welch's t-test since variances may be different)

```
> t.test(A,B)
        Welch Two Sample t-test
data:  A and B
t = -1.6073, df = 117.996, p-value = 0.1107
alternative hypothesis: true difference in means is
  not equal to 0
95 percent confidence interval:
 -2.494476  0.259348
sample estimates:
mean of x mean of y
 24.11748  25.23504
```

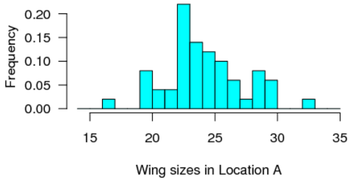
Kolmogorov-Smirnov (KS) test

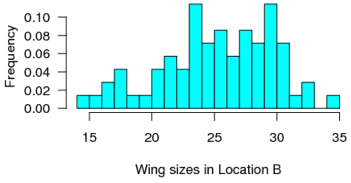
- So the means are not significantly different, what about the variances? Let's do Fisher's F-test

```
> var.test(A,B)
        F test to compare two variances
data:  A and B
F = 0.5014, num df = 49, denom df = 69, p-value =
  0.01192
alternative hypothesis: true ratio of variances is not
  equal to 1
95 percent confidence interval:
 0.3006728 0.8559914
sample estimates:
ratio of variances
         0.5014108
```

Kolmogorov-Smirnov (KS) test

- OK, the variances are different. Maybe also the skew? Let's have a look at the data anyway and plot histograms
  - Means are not significantly different
  - B has significantly larger variance





Kolmogorov-Smirnov (KS) test

- Kolmogorov-Smirnov test to compare the wing size distribution (location B) with the Normal distribution

```
> ks.test(B, "pnorm", mean=mean(B), sd=sqrt(var(B)))
        One-sample Kolmogorov-Smirnov test
data:  B
D = 0.0791, p-value = 0.7437
alternative hypothesis: two-sided
```

- Conclusion: the distribution of wing sizes in location B is not significantly different from the Normal distribution with the same mean and variance as the sample (p-value > 0.05)

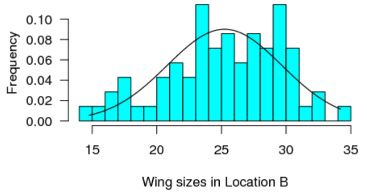
Kolmogorov-Smirnov (KS) test

- Since the Normal distribution can be used to describe the data in location B we can fit the Normal distribution to the data

```
> require(MASS)
> Bf <- fitdistr(B,"normal")
> Bfc <- coef(Bf)
> x <- seq(min(B),max(B),length=100)
> d <- dnorm(x, mean=Bfc[[1]], sd=Bfc[[2]])
> truehist(B, breaks=14:35,las=1, xlab="Wing sizes in
  Location B", ylab="Frequency")
> lines(x,d)
```

Kolmogorov-Smirnov (KS) test

- Here is the result: wing sizes in location B data as a histogram with the Normal distribution fitted to the data as a line



Exercise

- Test whether the wing sizes in location A are normally or log-normally distributed
  - Download wings.txt from WebCT and read into R
  - Split the wing size data into locations as described in the lecture
  - Test for normality
  - Test for log-normality
  - Fit the distribution(s)
  - Hint: log-transform your data and then test if the log-transformed data are normally distributed

Non-parametric versus parametric tests

- Parametric tests
  - make assumptions about how the data are distributed – which we usually do not know!
  - have a higher power (you will more likely be able to reject the null hypothesis when it is false)
- Non-parametric tests (distribution-free tests)
  - do not rely on assuming that data are normally distributed or that the data follow some other distribution
  - have less power
  - are conservative, they give the higher p-value, or less significant results
  - are more robust
  - Many non-parametric tests can also be applied to categorical data

Measurement scales

Quantitative or continuous variables	Ratio	Zero value defined so can calculate proportions	Concentrations, energy, lengths, absolute temperature Makes sense to say 'twice as much'
	Interval	Distances defined so can calculate differences	Temperature in Celsius, calendar dates
Categorical variables	Ordinal	Rank order defined	Low, medium, high (you can either rate or rank)
	Nominal	Distinctions defined	Bird, Mammal; Female, Male (you can classify but not rank)

Don't calculate sums, differences, means, etc.!

Non-parametric alternatives to the t-test

- Wilcoxon's rank sum test can be used even when the
  - data are not normally distributed
  - values are not quantitative (so you can only rank the values)
  - However, samples must be independent also for this test
- Principle of procedure
  - Values from both samples are pooled and this pooled list sorted and ranked
  - The ranks are summed separately for each sample (so you need to keep a sample label attached to your values for identification)
  - If the smaller of the rank sums is lower than the critical value, we reject the null hypothesis of equal location (location is the more general term as mean not valid for rank data)

Non-parametric alternatives to the t-test

- Example: comparing mean ozone concentrations of gardenA and gardenB

```
> wilcox.test(gardenA,gardenB)
        Wilcoxon rank sum test with continuity
  correction
data:  gardenA and gardenB
W = 11, p-value = 0.002988
alternative hypothesis: true location shift is not
  equal to 0
Warning message:
In wilcox.test.default(gardenA, gardenB) :
  cannot compute exact p-value with ties
```

Non-parametric alternatives to the t-test

- Comparing p-value with t-test shows that the parametric t-test has a higher power (though only if data are normally distributed, i.e. not skewed) and gives more significant results
  - t-test: p-value=0.001115
  - Wilcoxon's rank sum test: p-value=0.002988

Exercise

- Wilcoxon's rank sum test
  - Use wing size data as before
  - Is there a significant difference in mean wing sizes between location A and B?
  - Compare these results with t-test results

ANOVA

ANOVA

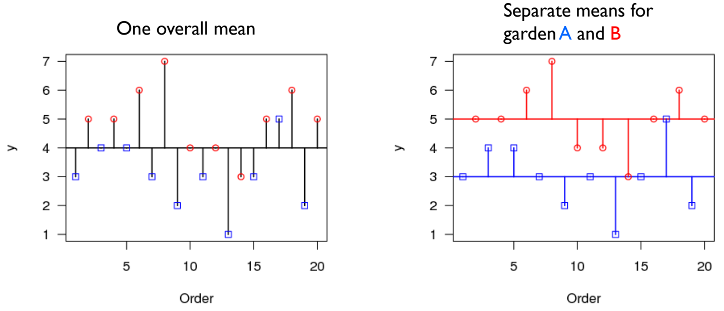
- Analysis of variance is used when your response variable is quantitative but all explanatory variables categorical
  - Regression is used when everything, your response variable and all explanatory variables, is quantitative (interval or ratio scale)
- It is called "analysis of variance" but is about comparing means!
  - Like the t-test, but can be used for > 2 samples/treatments
  - The key thing to understand is that we compare means by analysing variance
  - We partition variance into explained and unexplained parts, and test whether considering that different treatments have different means reduces our variance significantly

ANOVA

- Some terminology used in ANOVA
  - The explanatory variables are called factors (remember they are categorical)
    - Factors could be different types of treatments or conditions
    - Single factor: one-way ANOVA
    - Two factors: two-way ANOVA
    - Multiple factors: multi-way ANOVA
  - Each factor can come in two or more levels
    - If we had only one factor with two levels, e.g. only two treatments to compare, we could use the t-test
    - For example
      - Factor 1 with 3 levels: Drug with levels none, drug A, drug B
      - Factor 2 with 3 levels: Animal with levels mouse, rat, fish
  - When there is replication at each level for all factors, we have a factorial design

ANOVA

- When the means are significantly different, the sum of squares calculated from the individual treatment means will be smaller than the sum of squares calculated from the overall mean
  - This is the case in our example of ozone concentrations in our well known gardens A and B



ANOVA

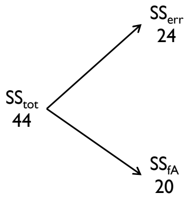
- ANOVA partitions variance
  - Variables: ozone (all), oa (ozone in garden A), ob (in garden B)
  - Total sum of squares $SS_{tot} = SS_{err} + SS_{fA}$
  - (using all data)

```
> SStot <- sum((ozone-mean(ozone))^2)
> SStot
[1] 44
```

ANOVA

- ‘Error’ (residuals) sum of squares SS_{err}

```
> SSerrA <- sum((oa-mean(oa))^2)
> SSerrA
[1] 12
> SSerrB <- sum((ob-mean(ob))^2)
> SSerrB
[1] 12
> SSerr <- SSerrA + SSerrB
> SSerr
[1] 24
```



- Treatment sum of squares SS_{fA} (treatment factor A)

```
> SSfa <- SStot - SSerr
> SSfa
[1] 20
```

ANOVA

- Now we can fill this information into our ANOVA table

	Sum of squares	Degrees of freedom	Variance = SS/d.f.	F-ratio
SS_{fA}	20	1 (Yes, only 1, since 1+18=19)	20	15
SS_{err}	24	18 (20 data points – 2 means)	24/18=1.33	
SS_{tot}	44	19 (20 data points – 1 mean)		

- We end up with an *F*-ratio
  - explained/unexplained variance = SS_{fA}/Ss_{err}
  - Test for significance in the usual way: is our F > critical value?

```
> qf(0.95,1,18)
[1] 4.413873
```

  - p-value in the usual way

```
> 1-pf(F,1,18)
[1] 0.001114539
```

ANOVA

- Let’s do a t-test

```
> t.test(oa,ob)
t = -3.873, df = 18, p-value = 0.001115
```

- Let’s do the whole ANOVA in one fell swoop

```
> model <- aov(ozone ~ garden)
> summary(model)
            Df  Sum Sq Mean Sq F value   Pr(>F)
garden       1 20.0000 20.0000      15 0.001115 **
Residuals   18 24.0000  1.3333

> summary.lm(model)
> plot(model, which=1:6)
```

ANOVA

- Comparison with t-test
  - ANOVA can compare more than two samples but if only 2, the results (p-values) are the same
  - ANOVA analyses variance by using F
    - F = explained/unexplained variance = SS_{fA}/Ss_{err}
  - t-test analyses differences by using t
    - t = difference between means / standard error of the difference
  - $F = t^2$

The Problem of Multiple Comparisons

The problem of multiple comparisons

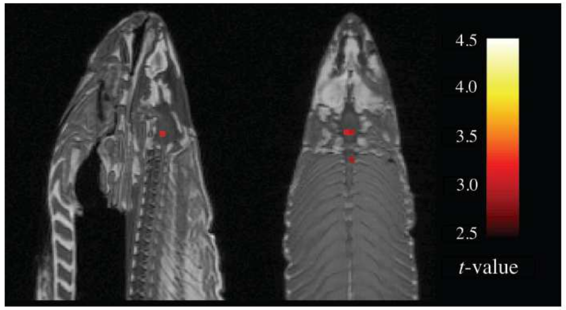
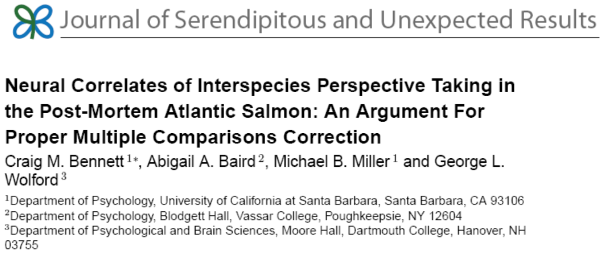


Fig. 1. Sagittal and axial images of significant brain voxels in the task > rest contrast. The parameters for this comparison were $t(131) > 3.15$, $p(\text{uncorrected}) < 0.001$, 3 voxel extent threshold. Two clusters were observed in the salmon central nervous system. One cluster was observed in the medial brain cavity and another was observed in the upper spinal column.

The problem of multiple comparisons



Journal of Serendipitous and Unexpected Results

Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction

Craig M. Bennett [1*], Abigail A. Baird [2], Michael B. Miller [1] and George L. Wolford [3]

[1] Department of Psychology, University of California at Santa Barbara, Santa Barbara, CA 93106
[2] Department of Psychology, Blodgett Hall, Vassar College, Poughkeepsie, NY 12604
[3] Department of Psychological and Brain Sciences, Moore Hall, Dartmouth College, Hanover, NH 03755

The problem of multiple comparisons

- If we do a single comparison, e.g. drug A effect on symptom 1 versus placebo, with standard significance level $\alpha = 0.05$
  - prob. for rejecting H_0 in error: 5%
- If we do 100 independent comparisons, e.g. drug A effect on independent symptoms S_1-S_{100}, it is highly likely that we reject H_0 for some symptoms even when H_0 is true in all 100 cases
  - Expected number of type I errors: 5
  - Probability of at least 1 type I error: 99.4%
    - prob. of not rejecting H_0 per single comparison $(1\text{-}\alpha) = 0.95$
    - prob. of not rejecting H_0 for all 100 (independent) comparisons: $(1\text{-}\alpha)^{100}$
    - prob. for rejecting at least one H_0 in error: $1\text{-}(1\text{-}\alpha)^{100} = 99.4\%$

The problem of multiple comparisons

- Multiple comparisons not just for ANOVA but common nowadays with high-throughput data
  - omics
  - microarrays: 10,000 or more comparisons simultaneously
  - fMRI and other imaging 100,000 or more

Solutions for multiple comparisons

- General case: comparisons may not be independent
  - Correct experiment-wise significance level (a.k.a. family-wise error rate) α by dividing by the number of comparisons n
  - Bonferroni correction $\alpha_{\text{(for single comparison)}} = \alpha\ /n$
    - But less power (higher type II error rate)
  - Holm-Bonferroni method more powerful
    - Order p-values, if smallest p-value < α /n, reject corresponding H_0 and remove this comparison from further analysis, then redo test with new Bonferroni correction α /(n-1) until remaining H_0 can no longer be rejected
- Extreme case: all comparisons entirely independent
  - Šidák correction $\alpha_{\text{(for single comparison)}} = 1\text{-}(1\text{-}\alpha)^{1/n}$
  - Advantage: gives higher $\alpha_{\text{(for single comparison)}}$ than Bonferroni correction
  - Tukey's HSD (Honestly Significant Difference) test for ANOVA results
- Extreme case: all comparisons entirely dependent
  - This is like doing only one comparison, so there is no need for corrections

ANOVA Exercise

- Read in data from "growth_burden.txt" into dataframe
  - Use read.table(), attach(), names(), as you learnt before
  - These data show the growth rates per min of different bacteria burdened by different plasmids, in different growth conditions
  - How many factors with how many levels?
  - Use str() to find out
  - For ANOVA, we have to force temperature to be a factor
    - Temp <- as.factor(Temp)
- Now do an ANOVA with aov() and access the results with summary() and summary.lm()
  - Which differences are significant?
- Use Tukey's Honestly Significant Difference test to do all pairwise comparisons with correcting for multiple comparisons
  - Which differences are significant?