UCLA STAT 13

Introduction to Statistical Methods for the Life and Health Sciences

•**Instructor:** Ivo Dinov,

Asst. Prof. In Statistics and Neurology

•**Teaching Assistants:** Tom Daula and Kaiding Zhu,

UCLA Statistics

University of California, Los Angeles, Fall 2002

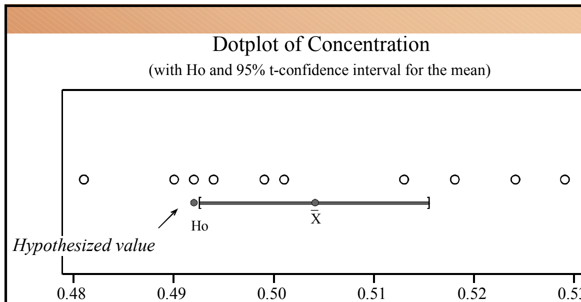
http://www.stat.ucla.edu/~dinov/

Chapter 10: Data on a Continuous Variable

- One-sample issues
- Two independent samples
- More than 2 samples
- Blocking, stratification and related samples

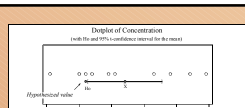
T-test and CI for the nitrate ion concentration data (mg/mL) in H_2O

- 10 samples measuring the NO_3^- ion concentration (possible fertilizer leak) in H_2O are given {0.513, 0.524, 0.529, 0.481, 0.492, 0.499, 0.518, 0.490, 0.494, 0.501}. Each sample measure is obtained by taking a sample of the H_2O and performing spectral chemical analysis. There's concern that there is a change from the desired nitrate concentration of 0.492.
- The data are plotted on the next slide, no reason to believe data is not coming form Normal distribution.



NO_3^- concentration in H_2O

Potential problems:
-small data size (10)
-Uncertainty about normal assumptions.
Approach, T-test, since it factors sample size (df) and is robust w.r.t. non-Normality.



Computer Analysis Output

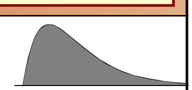
```
T-Test of the Mean                          se(x)        t0   P-value
Test of mu = 0.49200 vs mu not = 0.49200
Variable     N      Mean     StDev   SE Mean      T     P
Concentr    10   0.50410   0.01600   0.00506  2.390 040

T Confidence Intervals
Variable   N     Mean   StDev  SE Mean      95.0 % CI
concentr  10  0.50410 0.01600  0.00506 (0.49265, 0.51555
```


Robustness example

- Coverage of true mean by Normal-theory 95% CI's when sampling from the distribution depicted below
  - Coverage obtained by simulation
  - i.e. by repeatedly generating samples, calculating an interval from the sample and then determining whether the true value was in the interval or not.
  - This is called Chi-square distribution, χ^2(df=4), see in Ch. 11.
  - Results below represent 10,000 samples, each of 4 sample-sizes. Results are frequencies/percentage-of-time the t-interval covered the population mean. Skewed distribution, but good coverage.

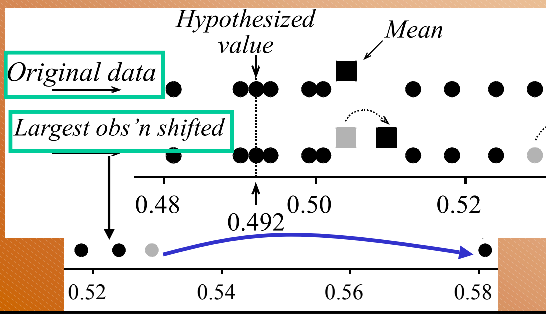
Sample Size	6	8	10	15
Coverage (%)	92	92	93	94



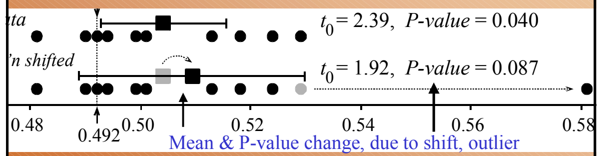
From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Effect of outliers

- We will shift one observation as shown below



Effect of an outlier



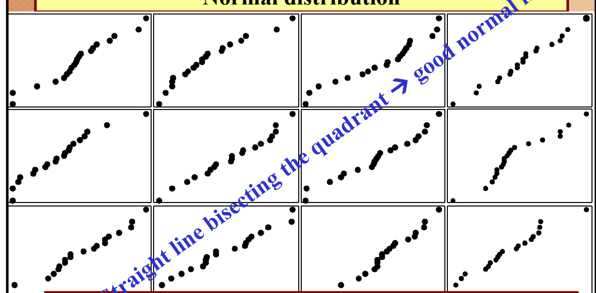
Data with largest obs'n shifted

```
Test of mu = 0.49200 vs mu not = 0.49200
Variable    N      Mean     StDev    SE Mean   t-stat.   P-val
Outlier    10   0.50930   0.02853    0.00902     1.92    0.087
                  95.0 % CI   (0.48889,0.52971)
```


Integrated approach to data analysis

Always **plot**, if you can, or **look** at your data **before using formal tools of analysis.**

Normal probability plots for samples of size 20 from a Normal distribution



- Normal probability/quantile plot allows a graphical test for normality of the data, by comparing its quantiles to these of a standard normal distribution. We use Normal samples, why are we getting non-normal fit? Small sample size!

Effect of outlier on Normal probability plot

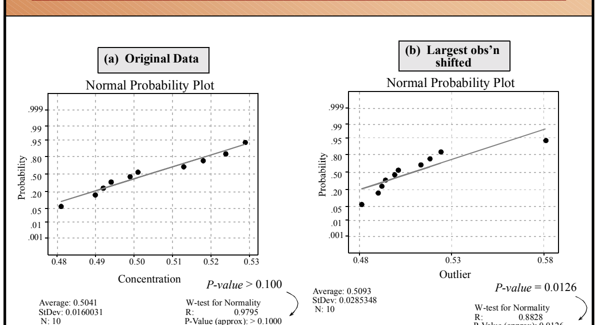


Figure 10.1.4 Normal plots for Example 10.1.1 data (from Minitab).

Example 10.1.2: Moon illusion

Does the Moon look larger at times? When its low in the sky close to the horizon, compared to the size being high in the sky. Starting with the Greek astronomer Ptolemy, 2nd century A.D., this problem has puzzled us until Kaufman and Rock described the illusion in 1962. They designed an experiment projecting two discs of adjustable size from the horizon direction (**level**) and from the **zenith** (directly overhead). Then compared the ratios of the disc diameters. One disc was kept at fixed size the other's size was updated until they appeared the same for the experimental subjects. Each measurement represents the ratio of the actual sizes zenith/horizon disc. If Moon-size-illusion occurs we'll have ratio>1, as the zenith disc size had to be increased to match the horizon disc size. {2.03, 1.65, 1.00, 1.25, 1.05, 1.02, 1.67, 1.86, 1.56, 1.73}

Example 10.1.2: Moon illusion

Data 2.03, 1.65, 1.00, 1.25, 1.05, 1.02, 1.67, 1.86, 1.56, 1.73

Assumptions: experimental subjects constitute random sample from large population. Hypothesis: H_0: $\mu=1$, H_a: $\mu>1$. One-sided P-value=0.0014. 95% CI(μ)=[1.21 : 1.75].

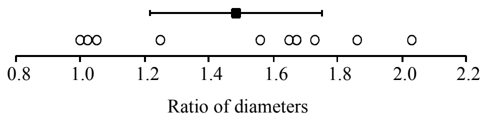


Figure 10.1.5 Dot plot of moon illusion data with 95% C.I. for mean

```
Test of mu = 1.000 vs mu > 1.000
Variable    N    Mean   StDev   SE Mean  t-stat  P-value
Elevated   10   1.482   0.374     0.118    4.07    0.001
                                  95.0 % CI:   (1.214, 1.750)
```


Comments

- What assumptions about the data are made by the theory underlying *t*-tests and confidence intervals for a population mean μ? (data are from distribution close to Normal)
- When we say that a *t*-confidence interval for μ is robust against some particular form of non-Normality, what do we mean by ***robust***? (Applies for non-normal data too, as long as there are no heavy outliers/clusters/skewed).
- What do we mean when we say that a *t*-test is it *robust* against some departure from the assumptions? (As the sample size increases we remove the requirement on the data as coming from Normal distribution, by CLT effects).

Review

- What should you always do with data on a continuous variable *before* performing formal significance tests or intervals? (Graph, Normal quantiles, eyeball).
- Under what circumstances should you ***not*** use *t*-tests and intervals?(small samples & skewed data–1-tailed test, outliers, clustered data).
- If there are outliers in a data set, what should you do? (check original data for typos, remove outliers)
- Four approaches to dealing with severe non-Normality (including the presence of outliers) are: non-parametric methods make no Normal assumptions (sign-test); robust methods insensitive to outliers; adopt a new model for the data underlying distribution (other than Normal) much like we did for T-distr; transformation approach (e.g., log-transform) to make the data conform better to Normal.

Paired Comparisons

- Sometimes we have two data sets, which are not independent, but rather observations matched in pairs.
- Back to the Kaufman & Rock study of the Moon size illusion. Does the moon size appear different with eyes level and with eyes raised? Does eye position make a difference? Eyes elevated refers to raising the eye from horizontal to zenith position. 10 Subjects are tested under eye-level (control) condition, by physically moving the subject's body from level to zenith position with fixed eye direction – horizontal. Ratios of the Moon size in level and zenith positions, for the two paradigms are given below.

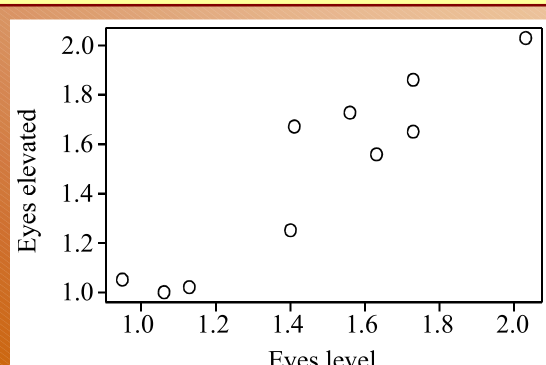
Moon illusion Data

TABLE 10.1.1 The Moon Illusion

Subject	Eyes Elevated	Eyes Level	Difference (Elevated - Level)
1	2.03	2.03	0.00
2	1.65	1.73	-0.08
3	1.00	1.06	-0.06
4	1.25	1.40	-0.15
5	1.05	0.95	0.10
6	1.02	1.13	-0.11
7	1.67	1.41	0.26
8	1.86	1.73	0.13
9	1.56	1.63	-0.07
10	1.73	1.56	0.17

Source: Kaufman and Rock [1962].

Plotting Eyes elevated rations vs. eyes level rations



Looking for an effect due to elevating eyes

For ***paired*** data, ***analyze the differences***.

$H_0: \mu_{diff} = 0$

Can't reject H_0, no evidence eye position causes illusion

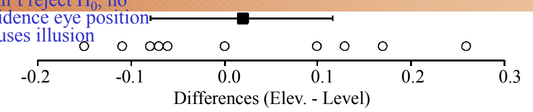


Figure 10.1.7 Dot plot of differences for the moon illusion data (with a 95% CI for the mean difference).

```
Test of mu = 0.0000 vs mu > 0.0000
Variable    N    Mean    StDev   SE Mean  t-stat  P-value
Difference  10   0.0190  0.1371  0.0434   0.44    0.34
                 95% CI ( -0.0791,  0.1171)
```


Flying helmet sizes for NZ Air Force

Measure the head-size of all air force recruits. Using cheaper cardboard or more expensive metal calipers. Are there systematic differences in the two measuring methods? Again, paired comparisons.

TABLE 10.1.2 Air Force Head Sizes Data

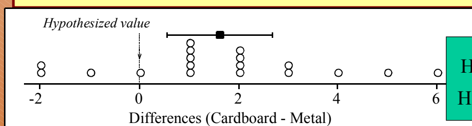
Recruit	Cardboard (mm)	Metal (mm)	Difference (Card-metal)	Sign of difference
1	146	145	1	+
2	151	153	-2	-
3	163	161	2	+
4	152	151	1	+
5	151	145	6	+
6	151	150	1	+

Helmet sizes for NZ Air Force – complete table

TABLE 10.1.2 Air Force Head Sizes Data

Recruit	Cardboard (mm)	Metal (mm)	Difference (Card-metal)	Sign of difference
1	146	145	1	+
2	151	153	-2	-
3	163	161	2	+
4	152	151	1	+
5	151	145	6	+
6	151	150	1	+
7	149	150	-1	-
8	166	163	3	+
9	149	147	2	+
10	155	154	1	+
11	155	150	5	+
12	156	156	0	0
13	162	161	1	+
14	150	152	-2	-
15	156	154	2	+
16	158	154	4	+
17	149	147	2	+
18	163	160	3	+

Head sizes: Does type of caliper make a difference?



$H_0: \mu_{diff} = 0$

$H_a: \mu_{diff} != 0$

Figure 10.1.8 Dot plot of differences in size (with 95% CI).

```
Paired T-Test and Confidence Interval
paired T for cardboard - metal

              N      Mean     StDev   SE Mean
cardboard     18    154.56     5.82     1.37
metal         18    152.94     5.54     1.30
Difference    18     1.611    2.146    0.506
95% CI for mean difference: (0.544, 2.678)
T-Test of mean difference=0 (vs not=0): T-Value=3.19
                                        P-Value=0.005
```

Figure 10.1.9 Minitab paired-*t* output for the size data.

Review

1. What is a paired-comparison experiment? (obs'd data are matched in pairs).
2. In a paired-comparison experiment, why is it wrong to treat the two sets of measurements as independent data sets? (data are usually taken from the same unit under diff. Treatments, so obs's should be related).
3. How do you analyze the data from a paired-comparison experiment? (analyze the difference).
4. What situations is appropriate to use the paired-comparison method to analyze the data? (pre- and post-metrifonate study using FDG PET imaging).

Helmet paired head measurements

From the cardboard vs. metal caliper tests, Table 10.1.2 we see 14 + and 3 – signs, implying larger overall measurements using the cardboard calipers. It's like tossing a coin 17 times and getting 14 heads. How likely is that?

If Y~Binomial(17, 0.5), number of successes (heads) in 17 fair coin tosses, then P(Y>=14)=0.00636, hence if we test p=0.5, vs. p!=0.5, two-tailed test, the chance is 2P(Y>=14)=0.0127.

Comments

5. What independence assumption must hold before the sign test is applicable? How important is it that this assumption is true? (requires that obs's are independent (one-sample test) and different pairs are independent (paired data), very sensitive.)

6. What advantages and disadvantages does the sign test have in comparison with the *t*-test? (Main advantage – test is distribution-free and insensitive to outliers. Disadvantage – when hypothesis for T-test, or a parametric test are met the CI are shorter and the parametric tests are more likely to detect departure from normality.)

Review

7. Why is the sign test called a distribution-free test? Does this mean that distributions are not used in performing the test? (no assumptions on the data underlying distribution, but distributions are actually used, e.g., Binomial).

8. In applying the sign test to paired data, how do you handle situations where both observations are tied (indistinguishable)? (ignore them)

Inference about the sample spread

Assessing the variability in the Moon Illusion Ratios. Three ratios in the range of the data are graphically shown below. Enormous variability in the extend of the illusion for different people. How to stat-analyze it.

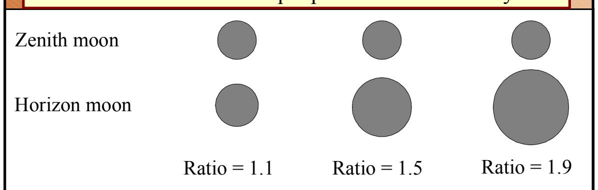


Figure 10.1.11 Different moon illusions.

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Inference about the sample spread

1. Here inferences for the mean μ are straightforward, whereas inferences for the standard deviation σ are difficult. (inferences about σ are very sensitive to non-normality, and does not improve with increase of sample-size.)

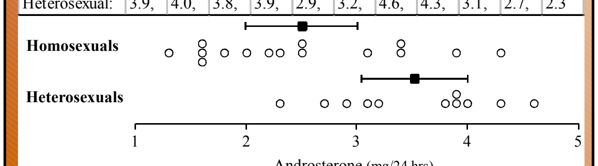
2. What large sample theory can we invoke when making inferences about a mean? (CLT, for μ, not available for σ)

Analysis of two independent samples

Urinary androsterone levels – data, dot-plots and 95% CI. Relations between hormonal levels and homosexuality, Margolese, 1970. Hormonal levels are lower for homosexuals. Samples are independent, as unrelated. Results, P-value of t-test 0.004 with a CI (μ_{Het}-μ_{Hom})=[0.4:1.7]. Normal hypothesis satisfied?Skewed?

TABLE 10.2.1 Urinary Androsterone Levels(mg/24 hr)

Homosexual:	2.5,	1.6,	3.9,	3.4,	2.3,	1.6,	2.5,	3.4,	1.6,	4.3,	2.0,
	1.8,	2.2,	3.1,	1.3							
Heterosexual:	3.9,	4.0,	3.8,	3.9,	2.9,	3.2,	4.6,	4.3,	3.1,	2.7,	2.3



Urinary androsterone levels cont.

Two Sample T-Test and Confidence Interval

```
Two sample T for androsterone
              N       Mean      StDev   SE Mean
hetero       11      3.518      0.721      0.22
homose       15      2.500      0.923      0.24
95% CI for mu (hetero) - mu (homose): ( 0.35,  1.69)
T-Test mu (hetero) = mu (homose) (vs not=):
                                        T=3.16  P=0.0044  DF=23
```

Confidence interval; *t-test statistic*; *P-value*

Figure 10.2.3 Minitab 2-sample *t*-output for the androstenone data

Comparing two means for independent samples

Suppose we have 2 samples/means/distributions as follows: $\{\bar{x}_1, N(\mu_1, \sigma_1)\}$ and $\{\bar{x}_2, N(\mu_2, \sigma_2)\}$. We've seen before that to make inference about $\mu_1 - \mu_2$ we can use a T-test for H_0: $\mu_1 - \mu_2 = 0$ with $t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{SE(\bar{x}_1 - \bar{x}_2)}$

And CI$(\mu_1 - \mu_2) = \bar{x}_1 - \bar{x}_2 \pm t \times SE(\bar{x}_1 - \bar{x}_2)$

If the 2 samples are independent we use the SE formula

$SE = \sqrt{s_1^2/n_1 + s_2^2/n_2}$ with $df = Min(n_1 - 1; n_2 - 1)$

This gives a conservative approach for hand calculation of an approximation to the what is known as the Welch procedure, which has a complicated exact formula.

Means for independent samples – equal or unequal variances?

Pooled T-test is used for samples with assumed equal variances. Under data Normal assumptions and equal variances of $(\bar{x}_1 - \bar{x}_2 - 0)/SE(\bar{x}_1 - \bar{x}_2)$, where

$$SE = s_p\sqrt{1/n_1 + 1/n_2};\ s_p^2 = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

is exactly Student's t distributed with $df = (n_1 + n_2 - 2)$

Here s_p is called the pooled estimate of the variance, since it pools info from the 2 samples to form a combined estimate of the single variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

The book recommends routine use of the Welch unequal variance method.

Comparing two means for independent samples

1. How sensitive is the two-sample *t*-test to non-Normality in the data? (The 2-sample T-tests and CI's are even more robust than the 1-sample tests, against non-Normality, particularly when the shapes of the 2 distributions are similar and $n_1 = n_2 = n$, even for small n, remember $df = n_1 + n_2 - 2$.
3. Are there nonparametric alternatives to the *two-sample t-test*? (Wilcoxon rank-sum-test, Mann-Witney test, equivalent tests, same P-values.)
4. What difference is there between the quantities tested and estimated by the two-sample *t*-procedures and the nonparametric equivalent? (Non-parametric tests are based on ordering, not size, of the data and hence use median, not mean, for the average. The equality of 2 means is tested and CI($\tilde{\mu}_1$ - $\tilde{\mu}_1$).

We know how to analyze 1 & 2 sample data. How about if we have than 2 samples – One-way ANOVA, *F*-test

One-way ANOVA refers to the situation of having one factor (or categorical variable) which defines group membership – e.g., comparing 4 reading methods, effects of different reading methods on reading comprehension, data: 50 – 13/14 y/o students tested.

Hypotheses for the one-way analysis-of-variance F-test

Null hypothesis: All of the underlying true means are identical.

Alternative: Differences exist between some of the true means.

Comparing 4 reading methods

Comparing 4 reading methods, effects of different reading methods on reading comprehension, data: 50 – 13/14 y/o students tested.

-Mapping: using diagrams to relate main points in text;

-Scanning: reading the intro and skimming for an overview before reading details;

-Mapping and Scanning;

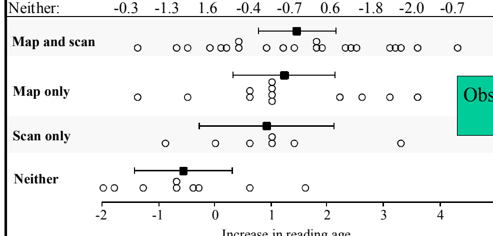
-Neither.

Table below shows increases in test scores, of 4 groups of students taking similar exams twice, w/ & w/o using a reading technique.

Research question: Are the results better for students using mapping, scanning or both?

TABLE 10.3.1 Increase in Reading Age

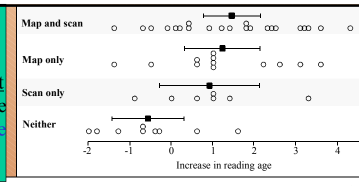
Both:	0.1	3.2	4.3	-0.5	1.9	3.3	2.5	3.6	0.4	2.3	-1.4	-0.7
	-0.1	0.2	0.4	0.9	1.2	1.4	1.8	1.8	2.4	3.1		
Map Only:	1.0	-0.5	1.0	0.6	0.6	1.0	1.0	-1.4	2.2	3.6	3.1	2.6
Scan Only:	1.0	3.3	1.4	-0.9	1.0	0.0	0.6					
Neither:	-0.3	-1.3	1.6	-0.4	-0.7	0.6	-1.8	-2.0	-0.7			



Observational study

Figure 10.3.1 Increases in reading ages with individual 95% CIs.

The F-test indicates that there's real evidence true differences exist it does not give indication of where the differences are or how large they are.




```
One-way Analysis of Variance
Analysis of Variance for Increase   F-statistic   P-value
Source   DF      SS      MS       F       P
Grp       3   27.06    9.02    4.45   0.008   } Anova Table
Error    46   93.35    2.03
Total    49  120.41
                             Individual 95% CIs For Mean
                             Based on Pooled StDev
Level       N    Mean   StDev  ------+---------+---------+---------+
MapScan    22   1.459   1.544                     (------*-----)
MapOnly    12   1.233   1.441                   (-------*--------)
ScanOnly    7   0.914   1.302              (----------*---------)
Neither     9  -0.556   1.135   (--------*---------)
                                ------+---------+---------+---------+
Pooled StDev =    1.425           -1.0       0.0       1.0       2.0
```

Figure 10.3.2 Minitab analysis of variance output for reading ages

Computer output


```
One-way Analysis of Variance
Analysis of Variance for Increase   F-statistic   P-value
Source   DF      SS      MS       F       P
Grp       3   27.06    9.02    4.45   0.008   } Anova Table
Error    46   93.35    2.03
Total    49  120.41
                             Individual 95% CIs For Mean
                             Based on Pooled StDev
Level       N    Mean   StDev  ------+---------+---------+---------+
MapScan    22   1.459   1.544                     (------*-----)
MapOnly    12   1.233   1.441                   (-------*--------)
ScanOnly    7   0.914   1.302              (----------*---------)
Neither     9  -0.556   1.135   (--------*---------)
                                ------+---------+---------+---------+
Pooled StDev =    1.425           -1.0       0.0       1.0       2.0
```

Figure 10.3.2 Minitab analysis of variance output for reading ages

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Interpreting the *P*-value from the *F*-test

(The null hypothesis is that all underlying true means are identical.)

- A ***large P-value*** indicates that the differences seen between the sample means could be explained simply in terms of sampling variation.
- A ***small P-value*** indicates evidence that real differences exist between **at least some** of the true means, but gives *no indication* of where the differences are or how big they are.
- ***To find out how big*** any differences are we need confidence intervals.

Form of a typical ANOVA table

TABLE 10.3.2 Typical Analysis-of-Variance Table for One-Way ANOVA

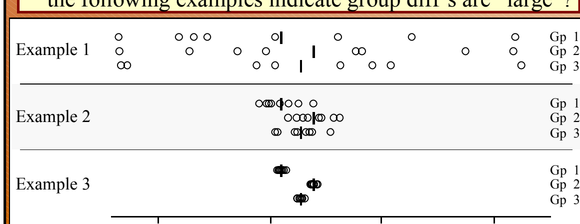
Source	Sum of squares	*df*	Mean sum of Squares[a]	*F*-statistic	*P*-value
Between	$\sum n_i(\bar{x}_{i\cdot}-\bar{x}_{\cdot\cdot})^2$	k - 1	s_B^2	$f_0 = s_B^2/s_W^2$	pr($F \geq f_0$)
Within	$\sum (n_i - 1)s_i^2$	n_{tot} - k	s_W^2		
Total	$\sum\sum (x_{ij}-\bar{x}_{\cdot\cdot})^2$	n_{tot} - 1			

[a]Mean sum of squares = (sum of squares)/*df*

- The *F*-test statistic, f_0, applies when we have independent samples each from ***k*** Normal populations, N(μ_i, σ), note same variance is assumed.

Where did the F-statistics came from?

- Let's look at this example comparing groups. How do we obtain intuitive evidence against H_0? Far separated sample means + differences of sample means are large compared to their internal (within) variability! Which of the following examples indicate group diff's are "large"?



More about the *F*-test

- s^2_B is a measure of variability of sample means, how far apart they are.
- s^2_W reflects the avg. internal Variability within the samples.

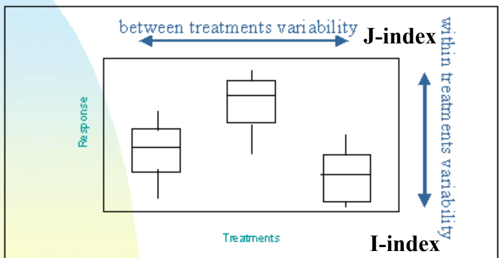
$$s_B^2 = \frac{\sum n_i(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2}{k-1}$$

$$s_W^2 = \frac{\sum (n_i - 1)s_i^2}{n_{tot} - k}$$

- The *F*-test statistic, f_0, tests H_0 by comparing the variability of the sample means (numerator) with the variability within the samples (denominator).
- Evidence against H_0 is provided by values of f_0 which would be unusually large if H_0 was true.

What are x_i, $x_{..}$, $x_{.j}$, etc.?

One-Way Anova (Sources of Variability)



What are x_i, $x_{..}$, $x_{.j}$, etc.?

Need Online reference

Apple juice sales (units per week)

H_0: $\mu_1 = \mu_2 = \mu_3$

H_A: at least 2 means differ

$x_{i,j}$, $1 \le i \le n_j$; $1 < j \le 3$

City 1 Convenience	City 2 Quality	City 3 Price
529	804	672
658	630	531
793	774	443
514	717	596
663	679	602
719	604	502
711	620	659
606	697	689
461	706	675
529	615	512
498	492	691
663	719	733
604	787	698
495	699	776
485	572	561
557	523	572
353	584	469
557	634	581
542	580	679
614	624	532

What are x_i, $x_{..}$, $x_{.j}$, etc.?

Sum of Squares for treatments (cities)

$$SST = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{\bar{x}})^2$$

$$SST = 20(577.55 - 613.07)^2$$
$$+ 20(653.00 - 613.07)^2$$
$$+ 20(608.65 - 613.07)^2$$
$$= 57{,}512.23$$

What are x_i, $x_{..}$, $x_{.j}$, etc.?

Sum of squares for the Error

Sum of Squares for Error: $SSE = \sum_{j=1}^{k} \left[\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 \right]$

$$SSE = 19(10{,}774.44) + 19(7{,}238.61) + 19(8{,}669.47)$$
$$= 506{,}967.88$$

What are x_i, $x_{..}$, $x_{.j}$, etc.?

F-test

Test Statistic: $F = \frac{MST}{MSE} = \frac{SST/(k-1)}{SSE/(n-k)}$

$$= \frac{57{,}512.23/(3-1)}{506{,}967.88/(60-3)}$$

$$= 3.23$$

Rejection Region: $F > F_{\alpha, k-1, n-k} = F_{.05, 2, 57} = 3.15$

Conclusion: Reject H_0

What are x_i, $x_{..}$, $x_{.j}$, etc.?

One-Way Design ANOVA Table

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F Statistic
Treatments	k-1	SST	MST	MST/MSE
Error	n-k	SSE	MSE	
Total	n-1	SS(Total)		

Note: MST=SST/(k-1)

MSE=SSE/(n-k)

F-test assumptions

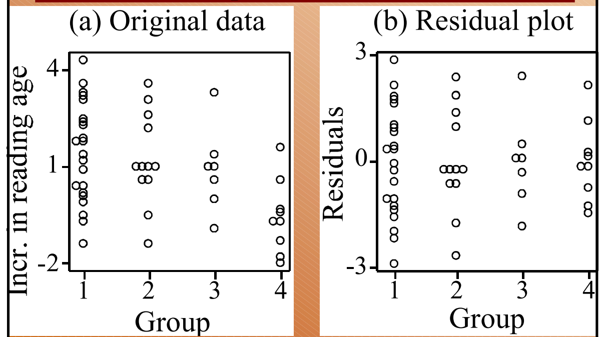
1. Samples are independent, physically independent subjects, units, objects are being studies.
2. Sample Normal distributions, especially sensitive for small n_i, number of observations, $N(\mu_i, \sigma)$.
3. Standard deviations should be equal within all samples, $\sigma_1 = \sigma_2 = \sigma_3 = \ldots \sigma_{n_k} = \sigma$. $(1/2 <= \sigma_k/\sigma_j <= 2)$

How to check/validate these assumptions for your data?

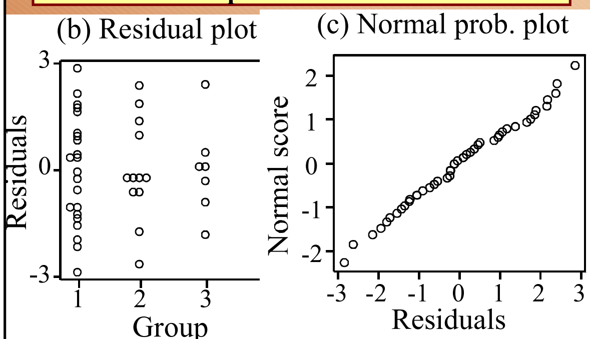
For the reading-score improvement data:

- independence is clear since different groups of students are used.
- Dot-plots of group data show no evidence of non-Normality.
- Sample SD's are very similar, hence we assume population SD's are similar.

Diagnostic plots for the reading-scores improvement data



Diagnostic plots for the reading-scores improvement data



Review

1. What is an one-way analysis of variance? (compare means of several groups of independent samples.)
2. When do we use the one-way ANOVA *F*-test? ($\{N(\mu_i, \sigma)\}_1^k$ samples).
3. What null hypothesis does it test? What is the alternative hypothesis? (all underlying true means are identical; at least 2 are different.)
4. Qualitatively, how does the *F*-test obtain evidence against H_0? (separation between sample means/intra-sample variability).
5. Qualitatively, what type of information is captured by the numerator of the *F*-statistic? What about the denominator? (variability-of-sample-means/variability-within-samples).

Review

6. Qualitatively, what values of f_0 provide evidence against H_0? (unusually large f_0 if H_0 is true.)
7. What does a large *P*-value from the *F*-test tell us about differences between means? How about a small *P*-value? (diff's between sample means can be explained by sampling variation.)
8. What does a small *P*-value tell us about which means differ from one another? about how big the differences between means are? (nothing about which/size, only indicates real diff's exist, between at least some sample means.)
9. How do we obtain information about the sizes of differences between means? (need confidence intervals.)

Review

10. What assumptions are made by the theory on which the *F*-test is based upon? How important is each of these assumptions in practice? (1.Sample independence – critical; 2.Normal data – robust, if sample-sizes are large; 3.Equal SD's – not too bad if $\sigma_{max}/\sigma_{min}<=2$.)
11. What new problem arises when we need to obtain and inspect a large set of confidence intervals? (all need to simultaneously catch, with 95% confidence, their true values, which requires increase of individual levels.)
12. Which is affected worst by departures from the equal-standard-deviations assumption, the *F*-test or the confidence intervals? Why? [CI, since CI(least-variable groups) = too wide & CI(most-variable-groups)=too narrow.]

Chapter 10 Summary

Always plot your data

Always plot your data before using formal tools of analysis (tests and confidence intervals).

- the quickest way to see what the data says
- often reveals interesting features that were not expected
- helps prevent inappropriate analyses and unfounded conclusions
- Plots also have a central role in checking up on the assumptions made by formal methods.

All formal methods make assumptions

- If the assumptions are false, the results of the analysis may be meaningless.
- A method is ***robust*** against a specific departure from an assumption if it still behaves in the desired way despite that assumption being violated.
  - e.g. it gives "95% confidence intervals" that still cover the true value of θ for close to 95% of samples taken.
- A method is ***sensitive*** to departures from an assumption if even a small departure from the assumption causes it to stop behaving in the desired way.

Assumptions cont.

- Many types of assumption are seldom, if ever, obeyed exactly so that methods which are sensitive to departures from such assumptions are of limited use in practical data analysis.
- You must check whether the data contradicts the assumptions to an extent where the tests and intervals no longer behave properly.
  - (Plots are a useful tool here.)

Outliers

- If present, try and check back the original sources.
- Any observations which you know to be mistakes should be corrected or removed.
- If in doubt, do the analysis with and without the outliers to see if you come to the "same" conclusions.

Nonparametric (distribution-free) methods

- less sensitive to outliers
- do not assume any particular distribution for the original observations
- do assume random samples from the populations of interest
- measure of center is the median rather than the mean
- tend to be somewhat <u>less effective</u> at detecting departures from a null hypothesis and tend to give <u>wider confidence intervals</u>

Normal Theory Techniques

One sample methods

- Two-sided t-tests and t-intervals for a single mean are
  - quite robust against non-Normality
  - can be sensitive to presence of outliers in small to moderate-sized samples
- One-sided tests are reasonably sensitive to skewness.
- Normality can be checked
  - graphically using Normal quantile plots
  - formally, e.g. the Wilk-Shapiro test.

Paired data

- We have to distinguish between independent and related samples because they require <u>different methods of analysis</u>.
- Paired data (Section 10.1.2) is an example of related data.
- With paired data, we analyze the differences
  - this converts the initial problem into a one-sample problem.
- The ***sign test*** and ***Wilcoxon rank-sum*** test are nonparametric <u>alternatives</u> to the one-sample or paired t-test.

2-sample t-tests and intervals for differences between means $\mu_1 - \mu_2$

Assume

- statistically independent random samples from the two populations of interest
  - both samples come from Normal distributions
- Pooled method also assumes that $\sigma_1 = \sigma_2$ Welch method (unpooled) does not

Two-sample t-methods are

- remarkably robust against non-Normality
- can be sensitive to the presence of outliers in small to moderate-sized samples
- One-sided tests are reasonably sensitive to skewness.

- The ***Wilcoxon*** or ***Mann-Whitney*** test is a nonparametric <u>alternative</u> to the two-sample t-test.

More than two samples and the F-test

- For testing whether <u>more than two means are different</u> we use the F-test.
- The method of comparing several means is referred to as a ***one-way analysis of variance***.
- The formal null hypothesis (H_0) tested is that all k ($k \geq 2$) underlying population means μ_i are identical.
- The alternative hypothesis (H_1) is that differences exist between at least some of the μ_i's.

The F-test cont.

- The numerator of the F-statistic f_0 reflects how far apart the sample means are. The denominator reflects average variability within the samples
- Evidence against H_0 is provided by
  - sample means that are further apart than expected from the internal variability of the samples.
  - large values of the F-statistic.
- A small P-value demonstrates evidence that differences exist between some of the true means
  - To estimate the size of any differences we use confidence intervals

Assumptions of the F-test cont.

- Assumptions of the F-test
  - independent samples;
  - Normality;
  - equal population standard deviations.
- The test
  - is robust to non-Normality
  - is reasonably robust to differences in the standard deviations when there are equal numbers in each sample, but not so robust if the sample sizes are unequal
  - can be used if the usual plots are satisfactory and the largest sample standard deviation is no larger than twice the smallest
  - is not robust to any dependence between the samples.