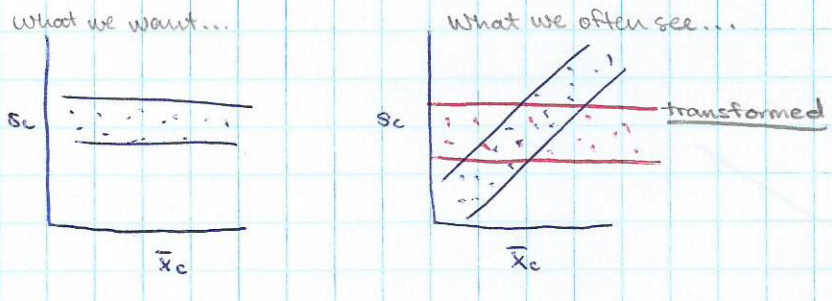
3/18/14

Exam 3 Material

Data Transformations

- Transformation = change in units of original data to better meet the assumptions of ANOVA or other test
- Use the transformed units in your tests (ANOVA)
- A general approach is to concentrate on equalizing the variances which often helps meet the other assumptions



- Need some kind of non-linear transformations

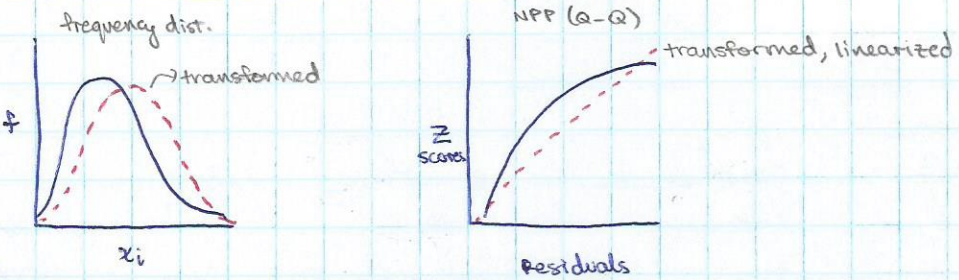
Popular Transformations

- Logarithmic (log or ln) transformation
- Log (ln) Transform

$$x_{ij}^* = \log(x_{ij} + 1)$$

→ by convention, add constant because you can't log 0

- Log transform can fix normality too



Square Root Transformation

$$x_{ij}^* = \sqrt{x_{ij} + 0.5}$$

again, want to avoid 0 and also neg values

→ by convention

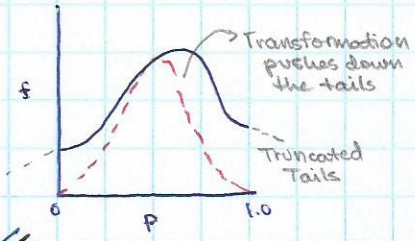
- Sqrt takes unequal variance and stabalizes it
- Also brings in extreme values to normal

3/15/2014

Arcsine Square Root transformation

- Useful for normalizing proportional or percentage data

$$p_{ij}^* = \arcsin\sqrt{p_{ij}}$$ p = proportion



Reciprocal Transformation

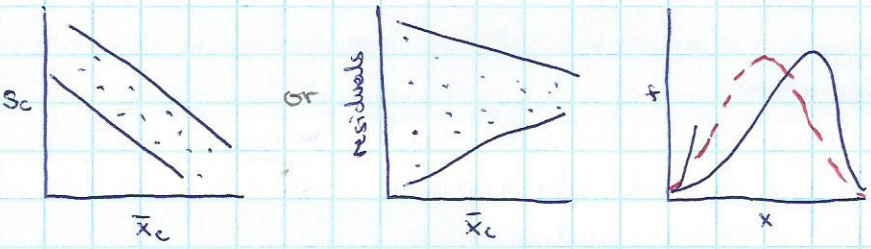
- Useful for very extreme values (for $s \propto \bar{x}^2$)

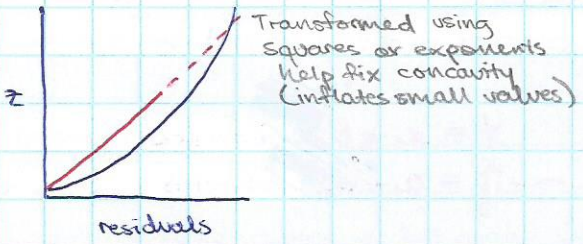
$$x_{ij}^* = \frac{1}{x_{ij}}$$

Square (or other exponent) Transformation

$$x_{ij}^* = x_{ij}^n$$

- Useful for rarer declining variance or SD or residual errors or left normal dist.





Exam Sections

1. 1 way ANOVA (F-test)
   a. balanced + unbalanced design (donuts)
   b. linear, additive models (FM, estimated)
   c. Multiple contrast test after sign. ANOVA
      i. Planned (a priori, LSD test)
      ii. Unplanned (a posteriori, HSD test)
   d. Power estimation $(1-\beta)$
   e. Model types - I, II, III
2. Randomized Blocks ANOVA
   a. remove extra variation (agri plots)
   b. antibiotic yields

3/18/2014

3. Multiway Factorial ANOVA
   a. ≥2 factors with ≥2 levels per factors (poisons + antidotes)
   b. factor interaction
      i. Synergistic
      ii. inhibitory
4. 2^n Factorial ANOVA (popcorn)
   a. each factor has 2 levels, 3 factors for pop
   b. Pondweed decay
5. ANOVA assumptions
   a. L.I.N.E.

3/25/2014

Linear Regression

Regression is a group of statistical methods used to examine functional relationships b/w variables

- Derived from ANOVA
- ANOVA asks "are means different?" and regr. asks "is there a trend in data?"
- Regression suggests a relationship exists but does not prove cause-effect

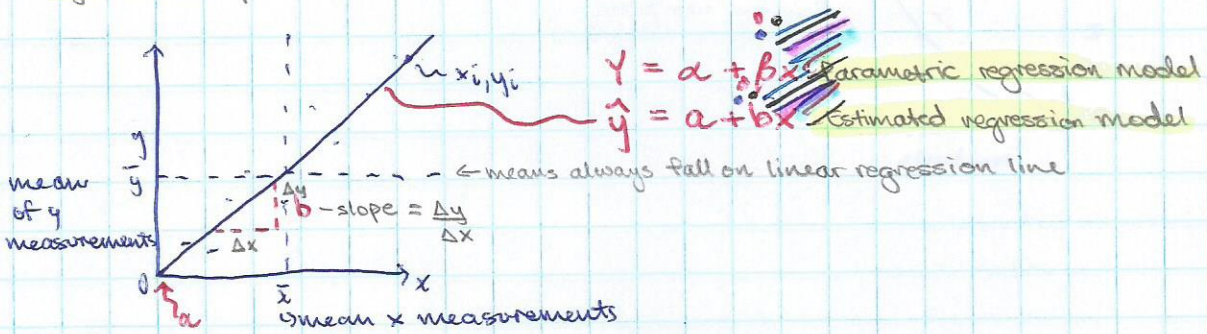
Simple Linear Regression

Simple Linear Regression describes relationship b/w two variables

Most basic SLR is unreplicated linear regression which is one observation of random variable y for every level of x

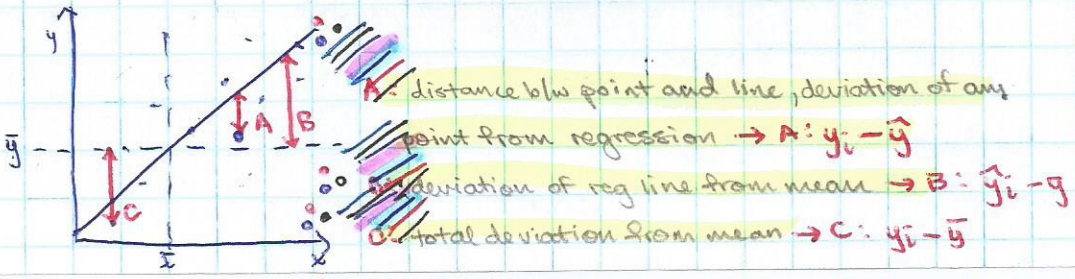
x is the indep. variable or predictor and is assumed to be known and fixed

y is the dependent variable or response and is measured with random error



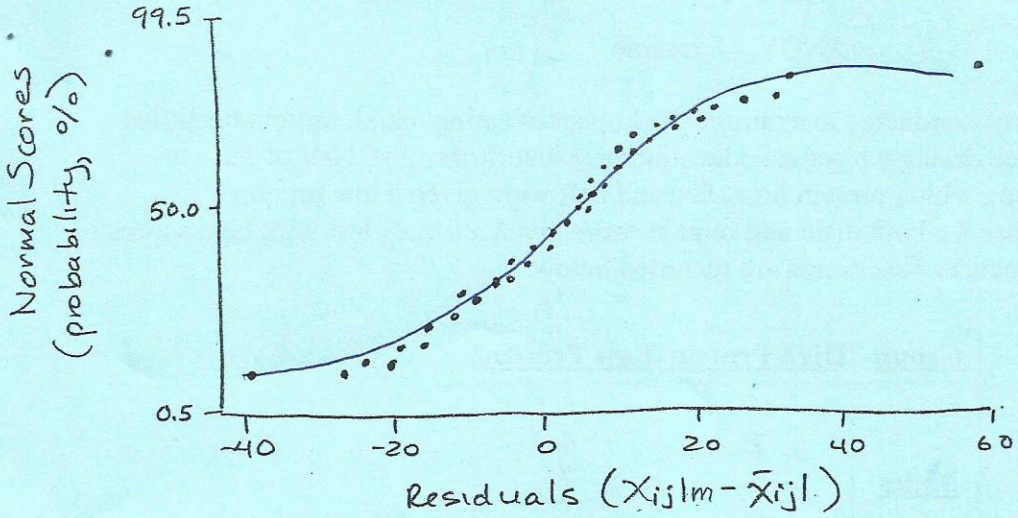
Relationship b/w x and y

We can say in regression: "value of y is dependent on or function of x value"

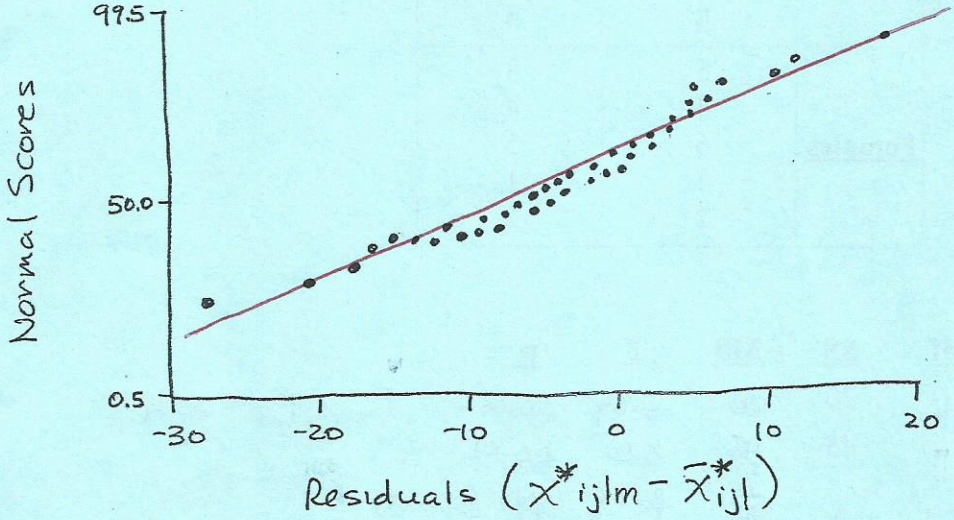


DATA TRANSFORMATION for 2^3 FACTORIAL ANOVA

3/18/2014



Normal Probability Plot of Residuals for original data (decay rates)



Normal Probability Plot of Residuals for Square-Root Transformed Data

ANOVA for original and transformed data:

Source	F, original data	F, transformed data
D	1.19 (ns)	1.44 (ns)
N	13.54 ($p<0.01$)	12.85 ($p<0.01$)
T	60.21 ($p<0.01$)	77.15 ($p<0.01$)
D×N	1.06 (ns)	1.69 (ns)
D×T	0.08 (ns)	0.39 (ns)
N×T	7.18 ($p<0.05$)	4.85 ($p<0.05$)
D×N×T	0.84 (ns)	2.32 (ns)

Rectifying non-normality does not greatly affect the outcome

ANOVA Example

A research investigation was conducted to examine the impact of eating a high protein breakfast on adolescents' performance during a physical education physical fitness test. Half of the subjects received, at random, a high protein breakfast and half were given a low protein breakfast. All of the adolescents, both male and female, were given a fitness test with high scores representing better performance. Test scores are recorded below.

Group	High Protein	Low Protein
Males	10	5
	7	4
	9	7
	6	4
	8	5
Females	5	3
	4	4
	6	5
	3	1
	2	2

Source	df	SS	MS	F	P
Protein Level	1	20	20	8.89	<0.01
Gender	1	45	45	20	<0.01
Protein Level x Gender	1	5	5	2.22	NS (>0.05)
Error	16	36	2.25		
Total					

$F_{0.05,1,4} = \approx 4.5$

Null Hypothesis(es):

Protein and gender have no effect on performance
No interaction b/w gender + protein

Conclusions:

Protein + Gender significant effect on performance
But no interaction between protein + gender

Was this a Model I, II, or III ANOVA?

Model III

Fixed protein level
Gender diffs are random

EXAM

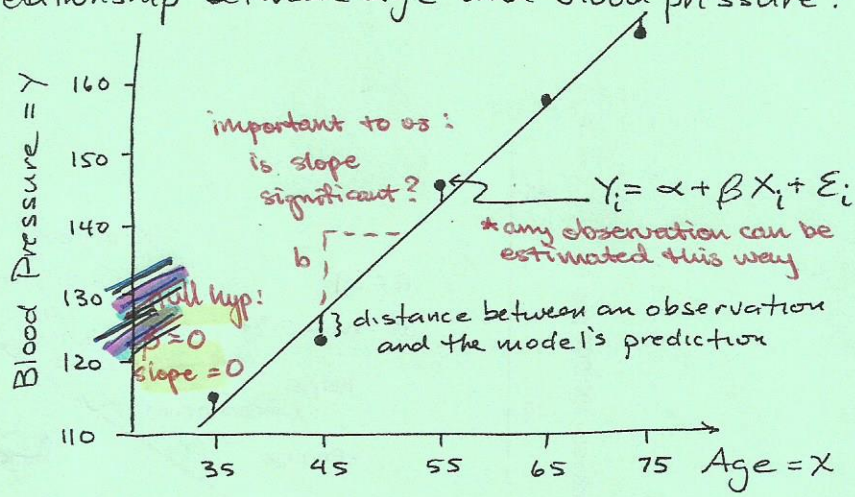
3x5, formulas, eq only, both sides
- all about ANOVA thru assumpt.
  - no transformations

LINEAR REGRESSION EXAMPLE

3/25/2014

Question: What is the functional relationship between age and blood pressure?

Age (x)	BP (y)
35	114
45	124
55	143
65	158
75	166
$\bar{x} = 55$	$\bar{y} = 141$



- ANOVA Table for Regression Results: F-test of $H_0: \beta = \emptyset$ — if it is not $\emptyset$ it is interesting

Source	df	SS	MS	F	Fcrit
Regression	1 (params − 1 (2−1))	$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	SS_R/df → MS_R	MS_R/MS_E	$v_1 = 1$, $v_2 = n-2$
Residual (Error)	$n-2$ (reps − number parameters estimated (a + b))	$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	SS_E/df → MS_E	—	
Total	$n-1$	$\sum_{i=1}^{n} (y_i - \bar{y})^2$	—		

$MS_E = s^2_{y \cdot x}$ ← variance of y on x

- For this Example

crit F = $F_{0.05,1,3} = 10.1$

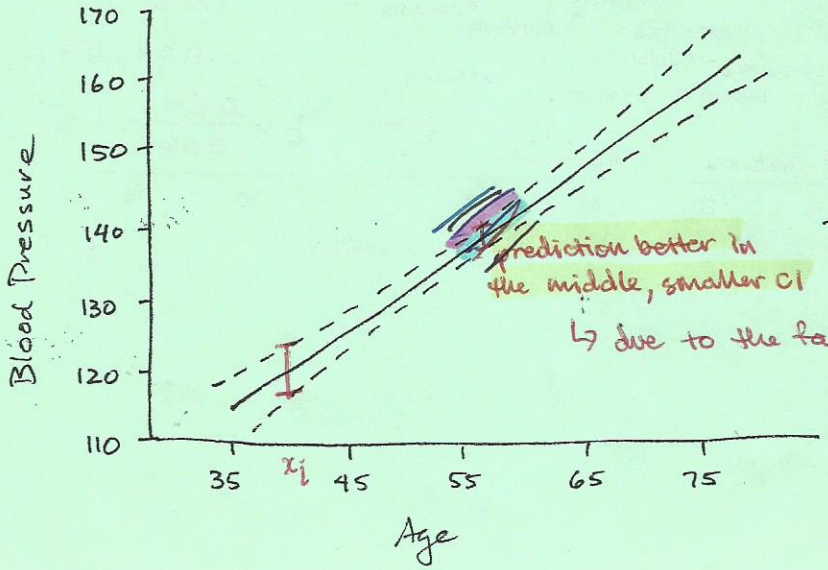
Source	df	SS	MS	F	P
Regression	1	1904.4	1904.4	180.9	<0.01
Residual	3	31.6	10.53		
Total	4	1936.0			

Reject H_0 that $\beta = 0$

↳ Slope is positive

Ratio of $\frac{SS_R}{SS_T} = R^2$

↳ Back at Notes

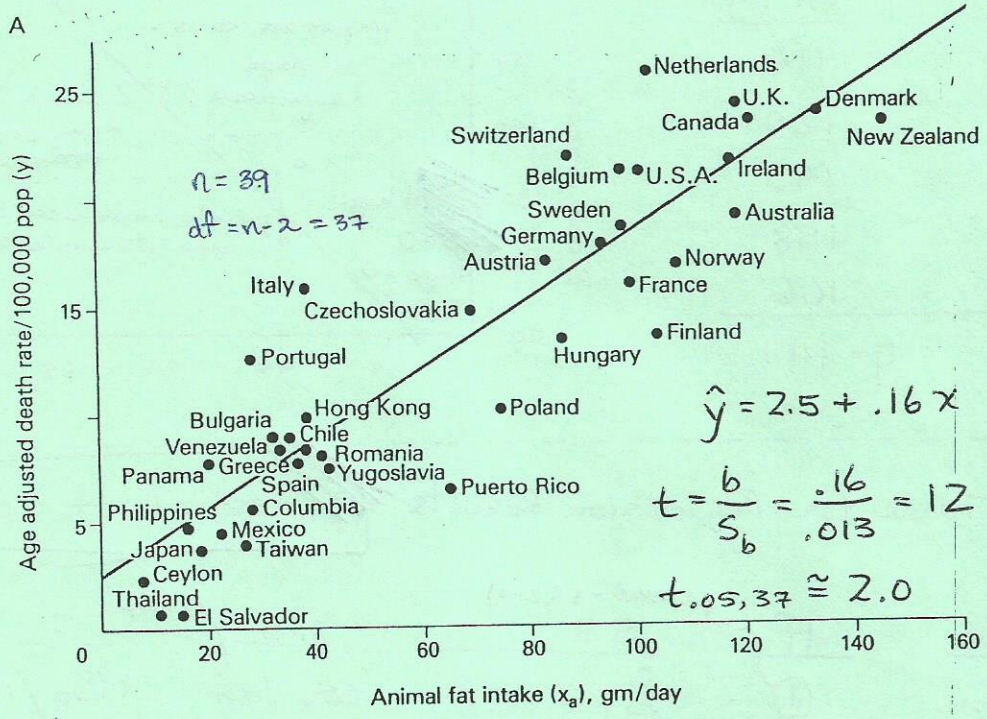


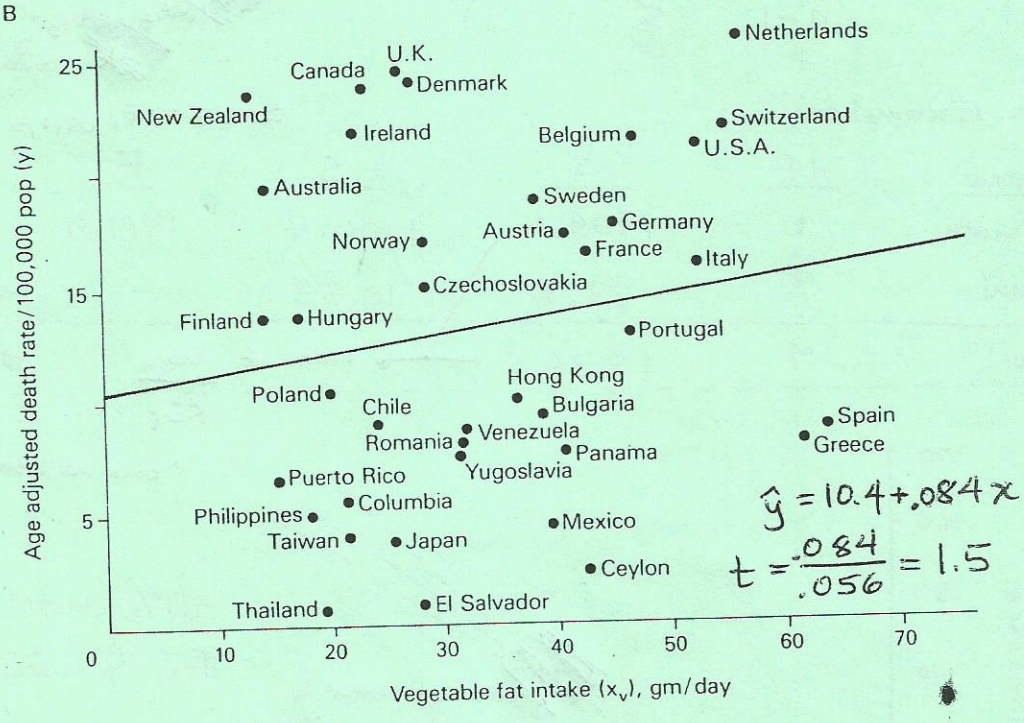
95% Confidence Belt for $\hat{y} = 65.1 + 1.38x$

prediction better in the middle, smaller CI

↳ due to the fact that $(x_i - \bar{x})^2$ in num. of $S_{\hat{y}_i}$ eq.

REGRESSION ANALYSIS USING T-DISTRIBUTION





3/25/2014

TOTAL SS = EXPLAINED SS + UNEXPLAINED SS

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \qquad n = \text{number obs.}$$

↳ TOTAL SS = REGRESSION SS + RESIDUAL SS

- The sums of squares gives us a means of quantifying variation in the model and summarizing it with an ANOVA table
  ↳ Green sheet (blood pressure and age)

- For any particular y response value, you can make parametric model, EM

PM: $Y_i = \alpha + \beta x_i + \varepsilon_i$

EM: $y_i = a + b x_i + (y_i - \hat{y}_i)$ ← how point deviates from line

- So how do we compute b & a?

$$b = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$ ← for any x_i, y_i (total num observations)
↳ called sum of cross products

= 1.38 for blood pressure example * Do at home...

$a = \bar{y} - b\bar{x}$ ← y-intercept

= 1.41 − (1.38)(55) for blood pressure

= 65.1

Regression Eq: for example: $\hat{y} = 65.1 + 1.38x$

- From SS_{regr} and SS_{total} in ANOVA table for BP example we calculate R^2 value

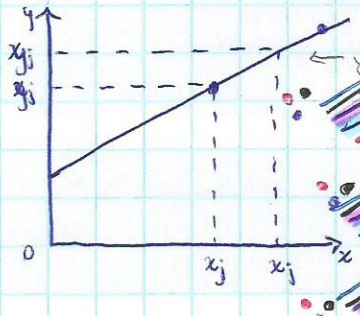
$$R^2 = \frac{SS_{reg}}{SS_{total}} = \frac{1904.4}{1936} = 0.984$$

↳ you would say 98.4% of variation in BP can be accounted for by age

↳ Also call R^2 the coefficient of determination

- you can also use regression equations for their predictive power

$\hat{y}_i = a + bx_j$ ← some other data point used to make prediction



← you can use equation to predict values at this point

↳ Error must be accounted for in prediction

$$S_{\hat{y}_i} = \sqrt{MS_E\left(\frac{1}{n} + \frac{(x_j-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}$$

↳ "Standard Error of prediction"

Confidence interval around prediction

$CL = \hat{y}_j \pm t_{\alpha, n-2} S_{\hat{y}_j}$ → CL around predicted y value

↳ take CLs and make confidence band/belt around regression line

3/27/2014

Relationship b/w x & y contd

- The regression equation minimizes deviations of the observations (y_i's) from the line, hence •least squares regression

Regression Warnings

- Totally unrelated variables should not be regressed, as results are misleading
  - ↳ Doesn't prove cause and effect
- However x can be random factor if levels cannot be specifically controlled (Model II)
- The SE of the prediction increases to the left or right of x
- When predicting y from x, interpolation is okay but extrapolation is not justified
  - ↳ keep predictions within data — XKCD extrapolation comic
- Confidence interval around prediction of regression eq.
  - ↳ a and b can have errors too!

$$S_b = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \quad \text{for BP ex.} \quad = \sqrt{\frac{10.53}{1000}} = 0.103 \quad \therefore b \pm s_b = 1.38 \pm 0.103$$

↳ SE of slope

$$S_a = \sqrt{MS_E\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)} \quad \text{For BP ex.} \quad = \sqrt{10.53\left(\frac{1}{5} + \frac{55^2}{1000}\right)} = 5.83 \quad \therefore a \pm s_a = 65.1 \pm 5.83$$

↳ SE of intercept

- Why go through error estimations?
  - ↳ Alternate means to test whether the slope β is $= \phi$
  - Instead of F-test, can do t-test where $H_0: \beta = \phi$

$$t = \frac{b-0}{s_b} = \frac{1.38}{0.103} = 13.45 \rightarrow t^2 = F = 180.9$$

↳ From ANOVA
↳ we can use t!

$$t_{crit} = t_{\alpha, n-2} = t_{0.05,3} = 3.182$$

$t > t_{crit}$ so reject H_0, $p < 0.001$

- Only for SLR, doesn't always work
- Sometimes we want to know if H_0: the $\alpha = 0$ (intercept thru origin)

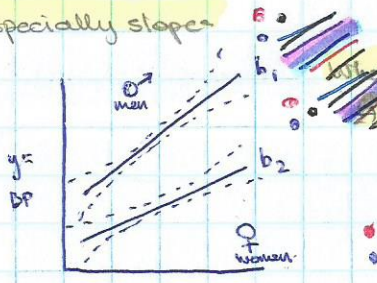
$$t = \frac{a-0}{s_a} = \frac{65.1}{5.83} = 11.17 \quad t_{crit} = 3.182 \quad t > t_{crit}$$

↳ Reject H_0

Comparing Regression Line

- Often we want to compare the coefficients of 2 regression lines

especially slopes



- What if we wanted to know $H_0: \beta_1 = \beta_2$
- Compare w/ t-test

$$t = \frac{b_1 - b_2}{s_{b_1 - b_2}} \qquad t_{\alpha, n_1+n_2-4}$$

↳ # of intercepts determined

↳ SE of Δ b/w slopes

3/27/14

$$s_{b_1-b_2} = \sqrt{\frac{(MS_E)_p}{\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2} + \frac{(MS_E)_p}{\sum_{i=1}^{n_2}(x_{2i}-\bar{x}_2)^2}}$$

$$(MS_E)_p = \frac{SS_{E1} + SS_{E2}}{df_{E1} + df_{E2}}$$

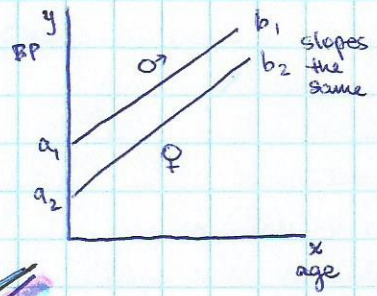
SS_E = residual sum of squares (from ANOVAs of regr.)

df = df of residual

but why not just use an ANOVA F-stat? Idk...

Similar to s_p^2 from independent t-test

- can do the same thing for intercepts



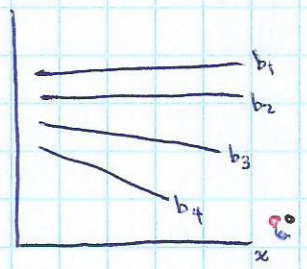
What if $H_0: a_1 = a_2$?

$$t = \frac{a_1 - a_2}{s_{a_1-a_2}} \sim t_{\alpha, n_1+n_2-4}$$

→ kind of complex to calculate, usually use software like R

We can use t-tests to compare 2 slopes, 2 intercepts

For 3 or more slopes, use analysis of covariance ANCOVA → will not discuss



Compare all slopes simultaneously $H_0: \beta_1 = \beta_2 \ldots \beta_k$

↳ just know we can do this, beyond scope of the class

- x is a covariate

If reject H_0, you can do MCTs

- Golden sheet, unreplicated simple linear regression

$$b = \frac{\sum_{i=1}^{n}(x_i-\bar{x}_i)(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad a = \bar{y} - b\bar{x}$$

$$b = [(0-50.39)(8.98-6.022)+(12-50.39)(8.14-6.022)+(29.5-50.39)(6.67-6.022)+(43-50.39)(6.08-6.022)$$
$$+(53-50.39)(5.90-6.022)+(62.5-50.39)(5.83-6.022)+(75.5-50.39)(4.68-6.022)$$
$$+(85-50.39)(4.2-6.022)+(93-50.39)(3.72-6.022)] / [(0-50.39)^2+(12-50.39)^2+(29.5-50.39)^2$$
$$+(43-50.39)^2+(53-50.39)^2+(62.5-50.39)^2+(75.5-50.39)^2+(85-50.39)^2+(93-50.39)^2]$$
$$= [(-50.39)(2.958)+(-38.39)(2.118)+(-20.89)(0.648)+(-7.39)(0.058)+(12.61)(-0.122)+$$
$$(12.11)(-0.192)+(25.11)(-1.342)+(34.61)(-1.822)+(42.61)(-2.302)] / [(2539.2)+(1473.8)$$
$$+(436.39)+(54.612)+(6.81)+(146.65)+(630.51)+(1197.9)+(1815.6)]$$
$$= [(-149.05)+(-81.31)+(-13.54)+(-0.43)+(0.32)+(-2.33)+(-33.6976)+(-63.06)+(-98.09)] / [(8301.47)]$$
$$= -0.053$$

$$a = 6.022 - (-0.053)(50.39) = 8.704$$

Reg Eq ⇒ $\hat{y} = 8.704 - 0.053x$

→

3/27/2014

2. ANOVA set-up, $H_0: \beta = 0$

	DF	SS	MS	F
Humidity	params − 1	$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	SS_H / df_H	MS_H / MS_E
Residuals	$n - 2$	$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	SS_E / df_E	

$SS_H = [(8.70-6.022)^2 + (8.07-6.022)^2 + (7.13-6.022)^2 + (6.42-6.022)^2 + (5.88-6.022)^2 + (5.38-6.022)^2 + (4.69-6.022)^2 + (4.18-6.022)^2 + (3.75-6.022)^2]$

$= [(7.17) + (4.19) + (1.23) + (0.16) + (0.02) + (0.037) + (1.77) + (3.393) + (5.162)]$

$= 23.132$

$SS_E = [(0.0784) + (0.0049) + (0.2116) + (0.1156) + (.0004) + (0.2025) + (0.0001) + (0.0004) + (0.0009)$

$= 0.6148$

	DF	SS	MS	F	
H	1	23.132	23.132	261	$F_{crit} = F_{0.05,1,7} = 5.59$
R	7	0.6148	0.09		$F > F_{crit} \rightarrow$ reject H_0, $p < 0.01$

3. $S_{\hat{y}_i} = \sqrt{MS_E \left(\frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}$ $\sum_{i=1}^{n} (x_i - \bar{x})^2 = 8301.47$ couple x predictions = 35, 65, 50.39

$S_{\hat{y}_i} = \sqrt{(0.09)\left(\frac{1}{9} + \frac{(35-50.39)^2}{8301.47}\right)} = 0.112 \rightarrow$ SE of pred. (35)

$S_{\hat{y}_i} = \sqrt{(0.09)\left(\frac{1}{9} + \frac{(65-50.39)^2}{8301.47}\right)} = 0.111 \rightarrow$ SE of pred (65)

$S_{\hat{y}_i} = \sqrt{(0.09)\left(\frac{1}{9}\right)} = 0.1$ SE of pred. (50.39) (middle)

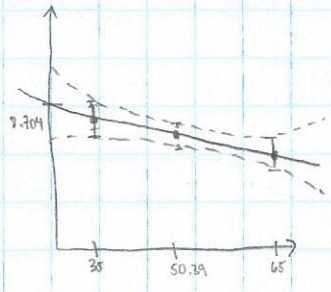
$\hat{y}_j = 8.704 - 0.053x$ $y(35) = 6.849$ $y(65) = 5.259$ $y(50.39) = 6.033$

$CL = \hat{y}_j \pm t_{\alpha, n-2} S_{\hat{y}_i}$

$CL(35) = 6.849 \pm (2.365)(0.112) = 6.849 \pm 0.265$

$CL(65) = 5.259 \pm (2.365)(0.111) = 5.259 \pm 0.263$

$CL(50.39) = 6.033 \pm (2.365)(0.1) = 6.033 \pm 0.237$



Something like that...

4. $R^2 = \frac{SS_{reg}}{SS_T}$ $SS_{Total} = SS_{reg} + SS_{resid} \rightarrow \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

3/27/2014

Lab 6. Acid rain data: Linear Regression
R output

TEMP v. SO2CONC

```
Call:
lm(formula = SO2CONC ~ TEMP, data = acid)

Residuals:
    Min      1Q  Median      3Q     Max
-33.247 -11.495  -1.982   3.543  71.187

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 117.3782a  26.8198Sa 4.377  9.1e-05 ***
TEMP         -1.5527b   0.4751Sb -3.268  0.0023 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.17 on 38 degrees of freedom
Multiple R-squared: 0.2194,     Adjusted R-squared: 0.1989
F-statistic: 10.68 on 1 and 38 DF,  p-value: 0.002300
```

↳ this is √ of (not negative of course)

$S_{y \cdot x}$

n = 40

ANOVA table:

```
            Df  Sum Sq Mean Sq F value   Pr(>F)
TEMP         1  4789.6  4789.6  10.683 0.002300 **
Residuals   38 17037.5   448.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

← MSE of $S^2_{y \cdot x}$

Shapiro-Wilk test

```
data:  resid(tempmod)
W = 0.8878, p-value = 0.000863
```

non-normal

1. Write regression equation, test $H_0: \beta = 0$ via ANOVA above, test $H_0: \alpha = 0$ via t-test, hand compute R^2

Reg Eq: $\hat{y} = 117.3782 + -1.5527x$

$H_0: \beta = 0 \rightarrow F_{stat} = 10.683 \rightarrow t = 3.268$ $t_{crit} = t_{0.05, 38} = 2.024$ $t > t_{crit}$ reject H_0

↳ Can reject based off of p value = 0.0023

$H_0: \alpha = 0 \rightarrow t = \frac{a - 0}{S_a} = \frac{117.3782}{26.8198} = 4.38$ $t_{crit} = 2.024$ $t > t_{crit}$, reject H_0

$R^2 = \frac{SS_{Reg}}{SS_T} = \frac{4789.6}{4789.6 + 17037.5} = 0.22$

↳ 22% of variation in SO_2 conc. can be accounted for by the data (verify w/ multiple R^2)

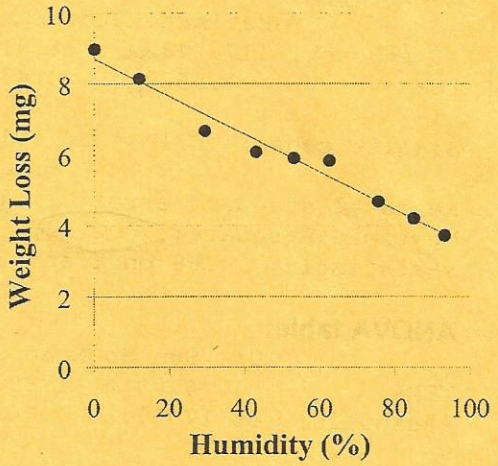
At home over the weekend

Unreplicated Simple Linear Regression

(do this at home)*

Evaluate weight loss (y, in mg) determined for 9 batches of 25 *Tribolium* flour beetles kept under differing humidity levels (x, as %). What is the functional relationship between humidity and weight loss?

Humidity (x, %)	Weight loss (y, mg)	$\hat{(y_i)}$
0.0	8.98	8.70
12.0	8.14	8.07
29.5	6.67	7.13
43.0	6.08	6.42
53.0	5.90	5.88
62.5	5.83	5.38
75.5	4.68	4.69
85.0	4.20	4.18
93.0	3.72	3.75
$\overline{x}$ = 50.39	$\overline{y}$ = 6.022	n = 9



1. Compute the regression coefficients (slope and intercept) and write the regression equation
2. Conduct the ANOVA Test of H_0: $\beta = 0$ and interpret it
3. Compute the standard error of some predictions (y_J), their confidence limits, and construct the confidence belt
4. Determine the coefficient of determination (R^2) and interpret it
5. Compute the standard errors of a and b
6. Conduct separate t-tests of H_0: $\beta = 0$ and H_0: $\alpha = 0$

**Solutions are on Sakai in Practice Problems folder*

4/1/2014

Simple Linear Regression Cont'd

SLR describes relationship b/w 2 variables

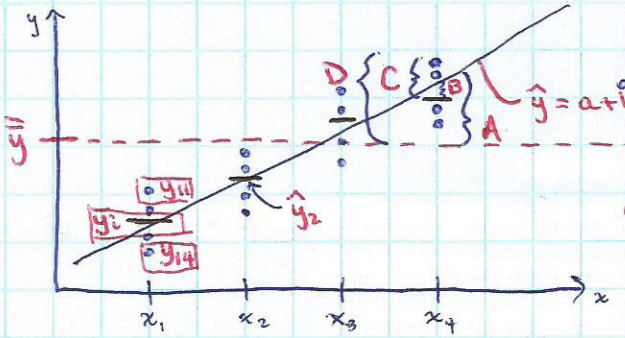
Recall that unreplicated linear regression is where we have one observation of random variable y for every level of x

↳ See blue handout

In replicated linear regression we have more than one y for each x which is more powerful than unreplicated SLR

↳ More predictable regression equation

Replicated Linear Regression



$$\text{Total SS} = \sum_{i=1}^{n_i}\sum_{j=1}^{n_j}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{n_i}\sum_{j=1}^{n_j}(\hat{y}_i - \bar{y})^2 + n_j\sum_{i=1}^{n_i}(\bar{y}_i - \hat{y}_i)^2 + \sum_{i=1}^{n_i}\sum_{j=1}^{n_j}(y_{ij} - \bar{y}_i)^2$$

(D = point − grand mean; A = reg − grand mean; B = mean − reg; C = point − mean)

EXPLAINED SS (REGRESSION) | LACK-OF-FIT SS | PURE ERROR SS

cancel, $(y_{ij} - \hat{y}_i)$

RE of residual in unreplicated SLR

$n = n_i n_j$ (total measurements)

$k = 4$ groups (four x values)

y_{ij} = j^{th} response, i^{th} level of x $\bar{y}_i$ = mean of all x_i measurements

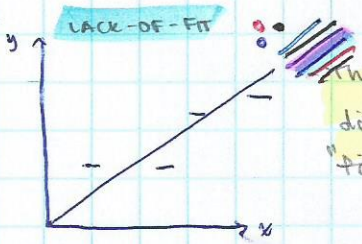
$\hat{y}_i$ = point on regression line that crosses x_i $\hat{y}$ = normal regr. equation

$\bar{\bar{y}}$ = grand mean A = difference b/w regression line and grand mean

B = deviation b/w group mean and regression

C = difference b/w mean and observations D = total (point − grand mean)

n_j = # of replicates per group n_i = # of levels of x (k)



This term describes how means differ from regression or don't "fit" to regression

Replicated Regression

Adds useful new feature – the ability to also test the null hypothesis that the linear model fits, versus only $H_0: \beta = 0$ in unrepl. regression

Based on deduction: if the model fits, then the variance of the cell means about the regression (lack of fit) should be due only to intrinsic variability in the data (pure error)

If deduction that linear is true:

$$F = \frac{SS_{LOF}/df_{LOF}}{SS_{PE}/df_{PE}} = \frac{MS_{LOF}}{MS_{PE}} = 1 \text{ if linear regression fits}$$

↳ want F value to be closer to 1

4/11/14

ANOVA Table

↗ params − 1 = 2 − 1 for LR

Source	DF	SS	MS	F	tests?
REGRESSION	1	$\sum_i n_i(\hat{y}_i - \bar{\bar{y}})^2$	SS_R/df_R	MS_R/MS_{EOF} ★	② $H_0: \beta = 0$ ↳ Different from ★ previous ANOVAs ↳ compare to LOF not PE
LACK OF FIT	# params − 2 = $k-2$	$\sum_i n_i(\bar{y}_i - \hat{y}_i)^2$	SS_{LOF}/df_{LOF}	MS_{LOF}/MS_E	① Is data linear? H_0: Linear ↳ compare to Fcrit ↳ If not linear, model doesn't work w/ data
PURE ERROR	$n-k$	$\sum_i \sum_j (y_{ij} - \bar{y}_i)^2$	SS_E/df_E		
TOTAL	$n-1$	$\sum_i \sum_j (y_{ij} - \bar{\bar{y}})^2$			

↳ Example on blue handout

Assumptions of Regression

1. Linear relationship exists b/w x and y
2. Values of y are obtained at random and are independent of each other
3. For any value of x, there is normal distribution of y's
4. For different values of x, variances of y's are equal

↳ Can log transform to fit some assumptions

↳ Test assumptions, transform if needed

↓

ANOVA for Rep. Regr.

- Linear Model Doesn't Fit → Revise Model → (back to ANOVA for Rep. Regr.)
- Linear Model Fits → $H_0: \beta = 0$ → R^2 and interpret

At Home Simple Replicated Linear Regr. Ex. (white sheet)

1. $SS_R = \sum_i n_i (\hat{y}_i - \bar{\bar{y}})^2$

$= [3(107.875 - 138.97)^2 + 4(120.905 - 133.97)^2 + 3(133.195 - 133.97)^2 + 5(146.965 - 133.97)^2 + 5(159.995 - 133.97)^2]$

$= [3(680.95) + 4(170.69) + 3(0.6) + 5(168.87) + 5(677.3)]$

$= (2042.85 + 682.76 + 1.8 + 844.35 + 3386.5)$

$= 6958.26$ close... maths on phone calc...

↳ 6750

At Home

4/1/2014

SIMPLE REPLICATED LINEAR REGRESSION: EXAMPLE*

Systolic Blood Pressure (y, mm Hg) collected for 20 individuals of different ages (x, years). What is the functional relationship between age and blood pressure?

Age (x, yrs)	Systolic BP (y, mm Hg)					n_i
$\bar{y}_{30} = 108$ 30	108	110	106			3
$\bar{y}_{40} = 120.5$ 40	125	120	118	119		4
$\bar{y}_{50} = 134.3$ 50	132	137	134		$\bar{\bar{y}} = 133.97$	3
$\bar{y}_{60} = 147.2$ 60	148	151	146	147	144	5
$\bar{y}_{70} = 159.8$ 70	162	156	164	158	159	5

Use the equation $\hat{y} = 68.785 + 1.303x$

$\hat{y}_{30} = 107.875$ $\hat{y}_{50} = 133.935$ $\hat{y}_{70} = 159.995$

$\hat{y}_{40} = 120.905$ $\hat{y}_{60} = 146.965$

1. Test for linearity (i.e., regression model fits the data) and H_0: $\beta = 0$

2. Calculate the Coefficient of Determination (R^2)

*Solution posted to Sakai, but make sure to work the problem fully first!

UNREPLICATED REGRESSION 4/1/2014

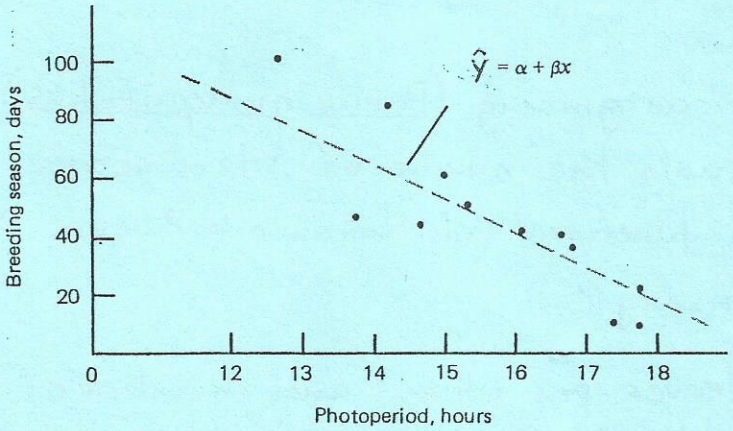


FIGURE 11.7
Theoretical line of regression and ideal curve for predicting the average length of the breeding season based on the photoperiod.

REPLICATED REGRESSION

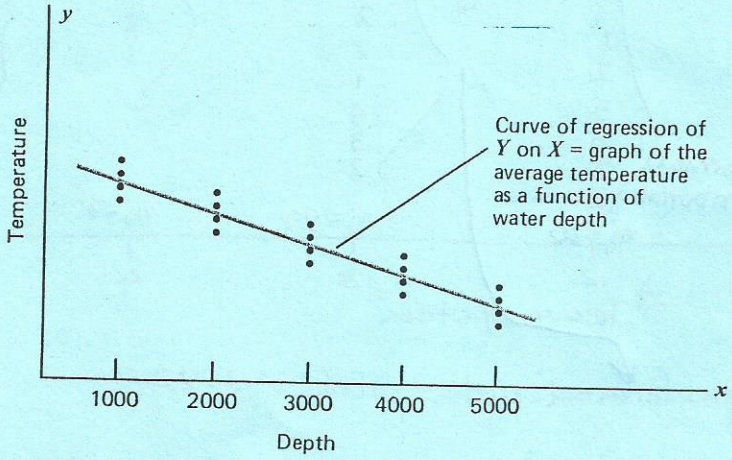


FIGURE 11.1
For a given water depth, the water temperature varies about some unknown average value The curve joining these mean values is called the curve of regression of Y on X.

REPLICATED LINEAR REGRESSION

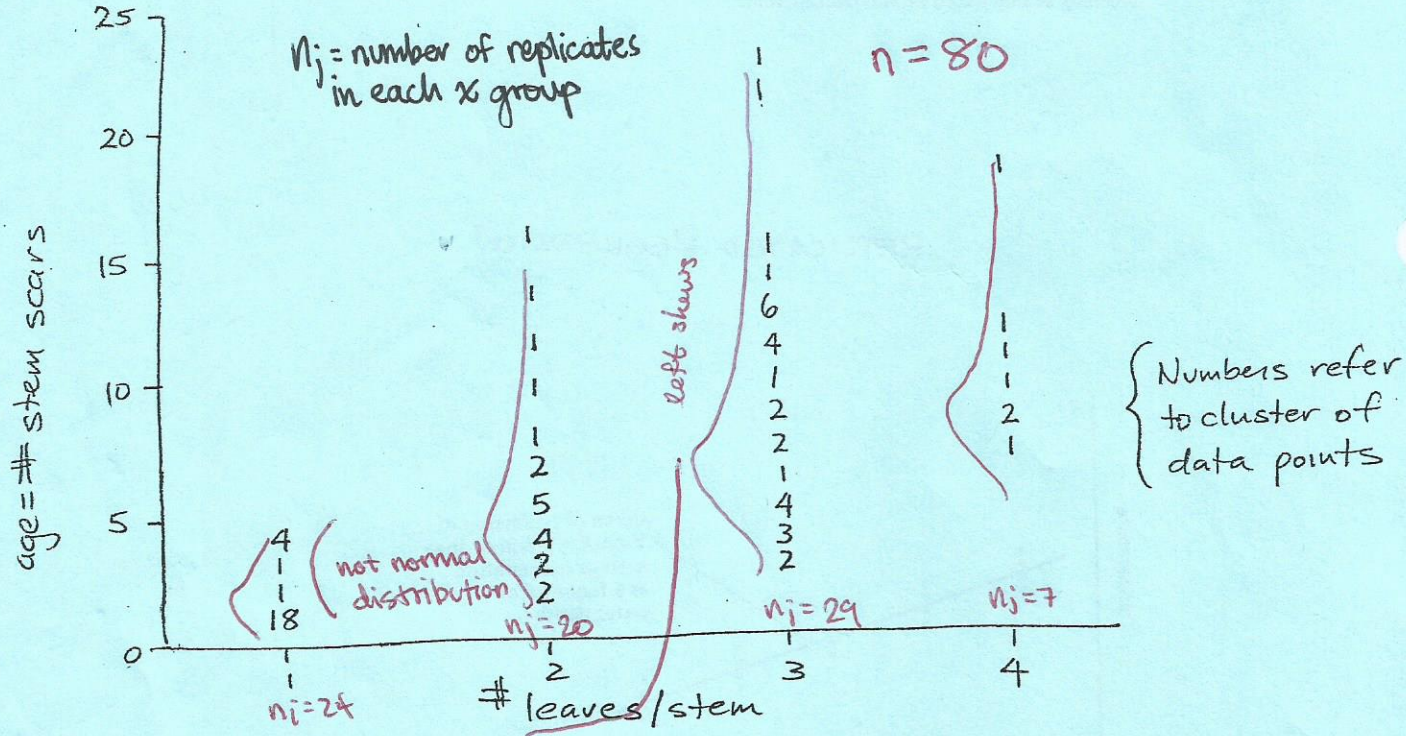
PARAMETRIC MODEL: $y_{ij} = \alpha + \beta x_i + L_i + \varepsilon_{ij}$ (LOF, ERROR)

ESTIMATED MODEL: $y_{ij} = a + bx_i + (\bar{y}_i - \hat{y}_i) + (y_{ij} - \bar{y}_i)$

where $\hat{y}_i = a + bx_i$

Problem: For American ginseng (*Panax quinquifolius*), plant age in years equals the number of stem scars on rhizome. Can we develop a functional relationship between # of leaves and age for ginseng?

Data: Number of leaves per stem and number of stem scars. Sample size is 80 plants.



Analysis of log-transformed data with ANOVA

Source	df	SS	MS	F	P
Regression	1	843.57	843.57	31.55	F_{crit} = 18.55, $0.025 > p > 0.05$, reject $b = 0$
Lack of Fit	2	53.47	26.735	1.88	$F_{crit} \sim 3.13$, $p > 0.05$, accept linear
Pure Error	76	1078.95	14.2	—	
Total	79		—		

$R^2 = \frac{SS_{REG}}{SS_{TOT}} = 0.43$

4/1/2014

$$SS_{LOF} = \sum_i n_j (\bar{y}_i - \hat{y}_i)^2$$

- $= [3(108 - 107.875)^2 + 4(120.5 - 120.905)^2 + 3(134.3 - 133.935)^2 + 5(147.2 - 146.965)^2 + 5(159.8 - 159.995)^2]$
  $= [(0.047) + (0.6561) + 0.3997 + 0.276 + 0.19]$
  $= 1.5689 \rightarrow 2$

$$SS_{PE} = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$$

$= [(108-108)^2 + (110-108)^2 + (106-108)^2] + [(125-120.5)^2 + (120-120.5)^2 + (118-120.5)^2 + (119-120.5)^2] + [(132-134.3)^2 + (137-134.3)^2 + (134-134.3)^2] + [(148-147.2)^2 + (151-147.2)^2 + (146-147.2)^2 + (147-147.2)^2 + (144-147.2)^2] + [(162-159.8)^2 + (156-159.8)^2 + (164-159.8)^2 + (158-159.8)^2 + (159-159.8)^2]$

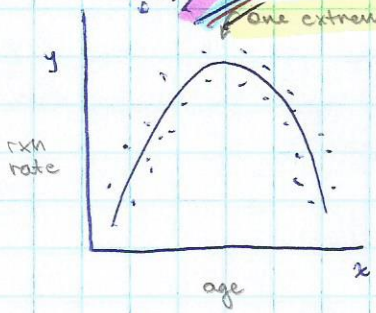
$= \sim 117$

Source	DF	SS	MS	F	p-value
Reg	1	6750	6750	3375	
LOF	3	2	0.667	0.0855	
PE	15	117	7.8		

4/3/2014

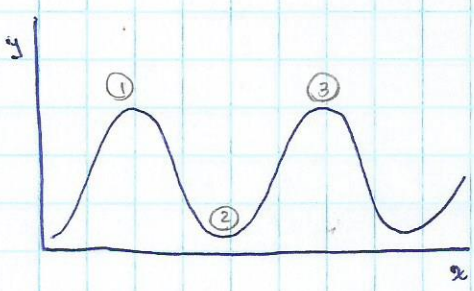
Non-Linear Regression

- Two popular types
  - Polynomial regression (curvilinear)
    - Basic Model:
      - POM: $Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \ldots \beta_n x_i^n + \varepsilon_i$
      - EM: $\hat{y}_i = a + b_1 x_i + b_2 x_i^2 + \ldots b_n x_i^n + (y_i - \hat{y}_i)$
    - For a parabola
      - one extreme



- $y = a + b_1 x + b_2 x^2$
  - ↳ quadratic equation or 2nd order polynomial
- ex. rxn rate vs. age
  - ↳ b's derived from multiple regression

Other polynomials: Oscillating systems



$y = a + b_1 x + b_2 x^2 + b_3 x^3 + b_4 x^4$

↳ 4th order polynomial

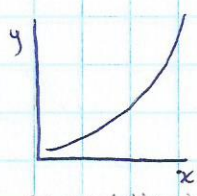
Polynomial Regression

- Polynomial of degree n-1 will fit n data points perfectly ($R^2 = 1$)
- When fitting more complex polynomials to data, need to address what they mean biologically
- In regression - simplest model possible
  - ↳ Relationship b/w x and y may mean nothing even w/ high R^2

Non-linear Regression cont'd

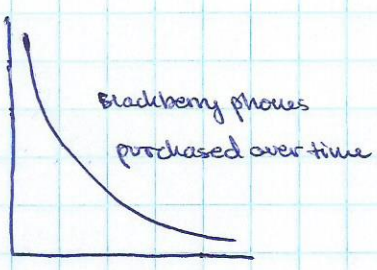
2. Exponential increase or decrease

Basic: $y = ae^{bx}$



ex. population increase

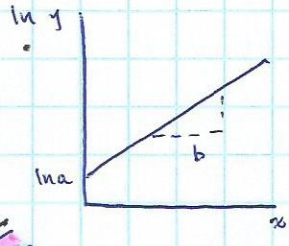
ex. radioactive decay

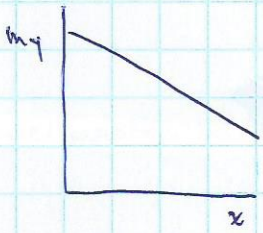


$\ln y = \ln a + bx$ Make that shit linear.

4/3/2014

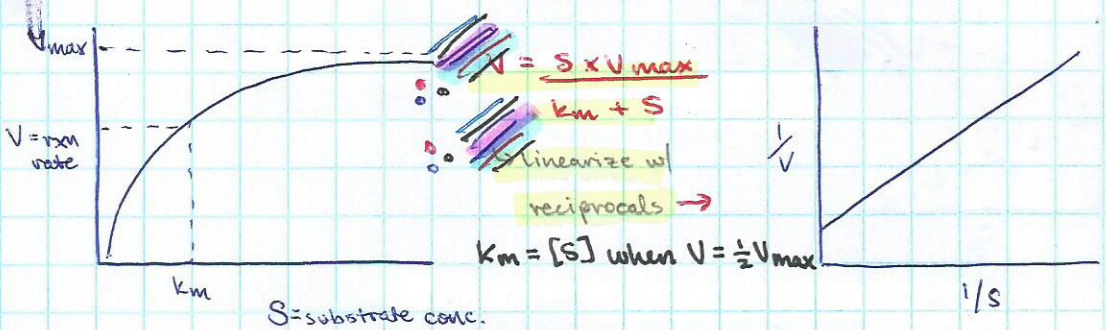
$\ln y = \ln a + bx$ linearize it! or $\ln y = \ln \cdot a - bx$





Michaelis Menten Equation

Enzyme kinetics mediated rxns

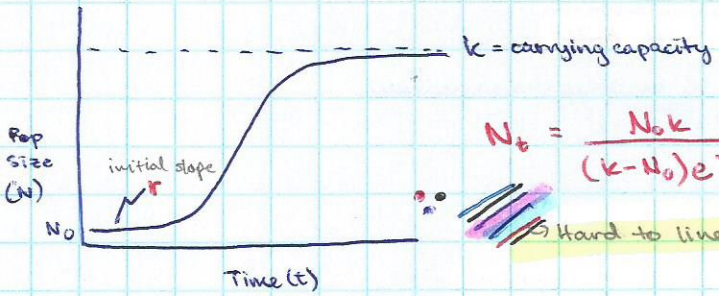


$V = \frac{S \times V_{max}}{K_m + S}$

linearize w/ reciprocals →

$K_m = [S]$ when $V = \frac{1}{2} V_{max}$

Logistic equation



$N_t = \frac{N_0 K}{(K - N_0)e^{-rt} + N_0}$

Hard to linearize... sqrt helps... ($\sqrt{N}$)

Special Problems w/ Non-linear Regression

- For eqs that can't be linearized, explicit equations for parameters (a, b, etc.) cannot be written, so params must be found by substitution
- Also sums of squares are not additive in this case so ANOVA and hypothesis tests are not possible
- If you can't make it linear, you can't do ANOVAs

Analysis of Residuals

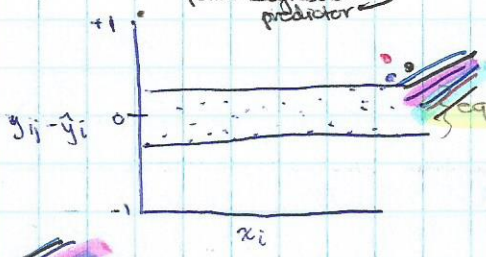
- As in ANOVA, examining regression residuals can help diagnose problems w/ the model or suggest transformation
- Recall that regression is an ANOVA so assumptions of ANOVA apply (L.I.N.E.)

4/3/2014

Types of Residual Plots

(1) Residuals $(y_{ij} - \hat{y}_i)$ vs predictors (x_i)

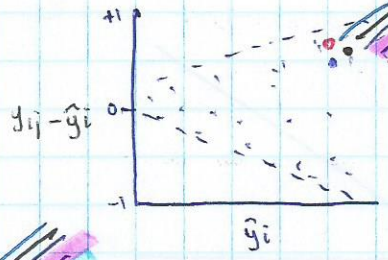
points — predictor



↳ equal variance (roughly)

(2) Residuals vs. predictions $(\hat{y}_i)$ (fitted values in R)

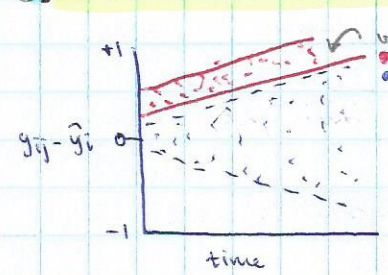
regression prediction



↳ increasing variance w/ increasing value $\hat{y}_i$

↳ log or sqrt to fix

(3) Residuals vs. order in which data were collected over time



↳ upward drift

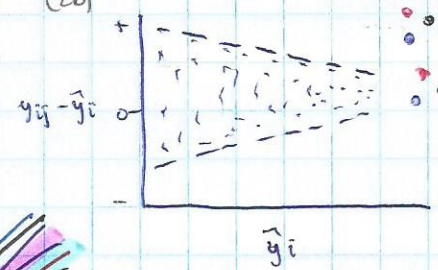
Ex: measurements on instrument

↳ observation decline in sensitivity

↳ or maybe drift

↳ fix by re-doing the experiment

(2b)



↳ increasing variance

↳ fix w/ n exponent

↳ Normal probability plots also useful

4/8/2014

Multiple Regression (Linear Regr)

↳ The purpose is to predict y from several variables (or x_i's)

↳ Example – atmosphere, humidity, windspeed to predict weather

↳ want to make sure the x's are actually related to y

↳ predictive variables should be rational

- Parametric Model

PM: $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} \dots \beta_n x_{ni} + \varepsilon_i$

EM: $y_i = a + b_1 x_{1i} + b_2 x_{2i} \dots b_n x_{ni} + (y_i - \hat{y}_i)$

4/8/2014

Remember the Null Hypotheses

- Overall $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_n = 0$
  - Test w/ ANOVA
- Per specific slope $H_0: \beta_i = 0$
  - Tested w/ specific t-test
  - A specific slope in MR tells you "how much y would change per unit change in x if the other x's were held constant"

$df_{Error} = n - (k+1)$ k = # of predictors (1 accounts for estimate of intercept)

$df_{model} = k$

For MLR, we are trying to select the best variables for the best model

Model Selection

- Often you will have many predictors (x_i) not all of which are needed to predict y
- High R^2 value is desirable
- As few variables as you need for accurate y_i prediction

Why minimize predictors

- Predictors may be costly or inefficient to measure
- Extraneous predictors make model too complex and unwieldy to use and interpret

Four Potential Approaches

- All possible regressions – use all combos of predictor variables

Ex. 3 variables x_1, x_2, x_3 → 7 equations

Possible Equations

$y_i = a + b_1x_1$ $\quad y_i = a + b_1x_1 + b_2x_2$ $\quad y_i = a + b_1x_1 + b_2x_2 + b_3x_3$

$y_i = a + b_2x_2$ $\quad y_i = a + b_2x_2 + b_3x_3$

$y_i = a + b_3x_3$ $\quad y_i = a + b_1x_i + b_3x_3$

- Advantage is you are considering all possible equations
- Disadvantage is as # of variables increase, number of equations get crazy

$$\#reg = \sum_{q=1}^{z} \frac{z!}{q!(z-q)!}$$

z = total # predictor variables

q = possible # predictor variables

Ex. $z = 10$

$$\frac{10!}{1!(10-1)!} + \frac{10!}{2!(10-2)!} + \dots \frac{10!}{10!(10-10)!} = 1023$$ yeesh...

- Stepping Backwards Elimination – eliminate predictor variables from model one by one
  - Usually better for smaller # predictive variables, less conservative α
  - Start w/ full model, remove insignificant variables
  - Time-saver, efficient and easy w/ R
  - Disadvantage – once you removed, variable is not examined again
    - Variable could be important in different context

4/8/2014

- Stepping Forward selection - add predictor variables not in model one by one
  - Keep adding significant variables until there are no more
  - Better for more variables
  - Start w/ simple model, more conservative α
  - Advantages include efficiency
  - Disadvantage is once you add a variable to the model, you can't remove it
- Stepwise Selection - add or eliminate predictors to or from model
  - Analyze the variables in and out of the model
  - Advantage over 2-3 because you can add back in or remove variables
  - In R, AIC values used to determine inclusion of variable in model
    - Smaller AIC value is better (- values good)
    - Can be less conservative and include insignificant variables

General Recommendations

Method	Approach	Advantage	Disadvantage
- All regressions	- Consider all combos of variables or predictors	- All possibilities considered	- Time consuming, hard to find best
- Backwards	- Start w/ all, eliminate one by one	- Efficient for a few variables	- Lose variables that might become important
- Forwards	- Start w/ simple model, add in one by one	- Efficient for a bunch of variables	- Include variables that become redundant
- Stepwise	- Can add or eliminate variables iteratively	- reflects importance of variables at each step	- Can include some arbitrary criteria by using AIC value

- For small # predictive variables (≤4): Do all possible regression
- Moderate # (>5): Backward or Stepwise
- Large # (>10): Forward or Stepwise

Criteria for Best Model

- Useful - can be interpreted and employed
- Parsimonious - as few predictors as possible
- Efficient - chosen predictor variables require minimal effort to acquire

Warning about MLR

- Replication becomes more important as number of variables increase
- Selected variables should be biologically relevant
- Large datasets often have multicollinearity
  - Some things are correlated

4/8/2014

Multiple Linear Regression

I. All Possible Regressions

1. Conduct all possible regressions with 1,2,3, … , z predictors and compare results
    - Advantage: All possibilities are examined.
    - Disadvantage: time consuming; no set criteria defining best model.

II. Stepping Backwards Elimination (BE)

1. Examine regression with all predictors (x_i) included.
2. Remove x_i with highest non-significant p-value for the test of H_0: $\beta_i = 0$.
3. Re-conduct multiple regression without selected x_i.
4. Continue removing predictors one at a time, recalculating H_0: $\beta_i = 0$ at each step, until all slopes are significant.
5. α can be considered more liberal at say 0.10.
    - Advantage: Efficient; all predictors are examined at once.
    - Disadvantage: Sometimes a x_i discarded from the model early in the process might become significant later in the analysis but BE does not consider significance of discarded x_i anymore.

III. Stepping Forward Selection

1. Examine H_0: $\beta_i = 0$ based on the minimum model $y_i = a$.
2. Select x_i with the lowest significant p-value and add to minimum model
3. Examine H_0: $\beta_i = 0$ based on the model $y_i = a + b_i x_{1i}$ (x_1= selected in step 2).
4. Continue adding predictors one at a time, recalculating H_0: $\beta_i = 0$ at each step and updating the model until all slopes outside the model are non-significant. Then run the multiple linear regression with the selected predictors.
5. α should be more conservative at 0.05 or less.
    - Advantage: Efficient, especially if there are many predictors and only a few are significant.
    - Disadvantage: Sometimes a x_i added to the model early in the process might become non-significant later in the analysis but FS does not consider significance of added x_i anymore.

IV. Stepwise

1. Examine H_0: $\beta_i = 0$ of all factors both within and outside the model at each step.
2. Add or remove predictors at each step based on significance (e.g., slope).
3. Continue until all slopes in model are significant and all slopes outside of model are non-significant (or selected model has the lowest AIC).
    - Advantage: Similar to BE and FS; also accounts for all predictors at every step thus countering the disadvantages of BE and FS.
    - Disadvantage: uses criteria some consider arbitrary (e.g., lowest AIC).

Call:

lm(formula = Ozone ~ Solar.R + Temp + Wind)

Residuals:

Min	1Q	Median	3Q	Max
-32.216	-13.922	-1.615	13.322	87.502

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-103.9449	51.3811	-2.023	0.049173 *
Solar.R	0.1057	0.0434	2.435	0.019009 *
Temp	2.1493	0.5467	3.932	0.000295 ***
Wind	-4.3541	1.1567	-3.764	0.000491 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.1 on 44 degrees of freedom

Multiple R-squared: 0.6448, Adjusted R-squared: 0.6206

F-statistic: 26.63 on 3 and 44 DF, p-value: 5.625e-10

Final Model :

$$y = -103.9449 + 0.1057 x_{solar.R} + 2.1493 x_{Temp} + -4.3541 x_{wind}$$

4/10/2014

Correlation

Correlation concerns the amount of association b/w two random variables

Correlation and regression not the same, even though the math is similar

How do they differ?

In regression we describe dependence of y on fixed levels of x

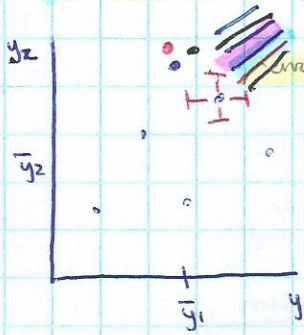
→ only y is measured w/ random error

→ Causation is suggested but not proven by regression

In correlation we want to know if 2 variables Y_1 & Y_2 are related or covary (vary with each other)

→ Both variables measured w/ error

→ Causation is not even implied w/ correlation



→ error in both directions

Bivariate Points (y_{1i}, y_{2i}) – random subjects from which we are measuring two variables

Ex.

Random Subjects	y_1	y_2
Humans	Cholesterol	Blood Pressure

→ In regression, cholesterol might be fixed x, with by varying y

Plants	Light	Photosynthesis
Years	Sunspots	Volcanic Activity
Vineyards	grape harvest	frog density

→ You can correllate whatever you want but some things probably won't be related

Correlation coefficients

Most popular measure is Pearson's r

$$r_p = r_{y_1 y_2} = \frac{\sum_{i=1}^{n}(y_{1i} - \bar{y}_1)(y_{2i} - \bar{y}_2)}{\sqrt{\sum(y_{1i} - \bar{y}_1)^2}\sqrt{\sum(y_{2i} - \bar{y}_2)^2}}$$

+pve or -nve n = # y_1, y_2 pairs, # observations, points

$$r_p = \frac{S_{y_1 y_2}}{S_{y_1} S_{y_2}}$$

$S_{y_1 y_2}$ = covariance of y_1 and y_2 (+ or −)

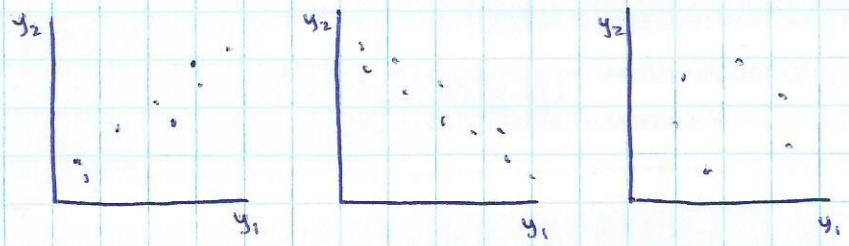
S_{y_1}, S_{y_2} = always +

↓ estimators (sample)

"rho" $\rho_p = \frac{\sigma_{y_1 y_2}}{\sigma_{y_1}\sigma_{y_2}}$ ← parameters (pop)

4/10/2014

- r can range from -1 to +1 and has no units
  - +1 means perfect positive correlation
  - -1 means perfect negative correlation
  - ϕ means no association



- r close to +1 — r close to -1 — r close to ϕ

- No regression lines!
- r is a single unitless # for the "linear" association b/w two variables
  - the "strength" of the relationship b/w Y_1 and Y_2
  - But can use regression to figure out r

$$r = b\frac{s_{y_1}}{s_{y_2}} \qquad r^2_{corr} = \frac{\sum(\hat{y}_{2i} - \bar{y}_2)^2}{\sum(y_{2i} - \bar{y}_2)^2} = \frac{SS_{REG}}{SS_{TOTAL}} = R^2_{regr.}$$

Correlation ← → Regression

r $\qquad$ R^2

Coefficient of Determination (r^2)

- r^2 = proportion of variation in one variable explained by variation in the other variable
- But in correlation we are most interested in r not r^2
- In regression we are very interested in R^2 the coefficient of determination
  - the capital R distinguishes from correlation r

Significance Tests in Correlation

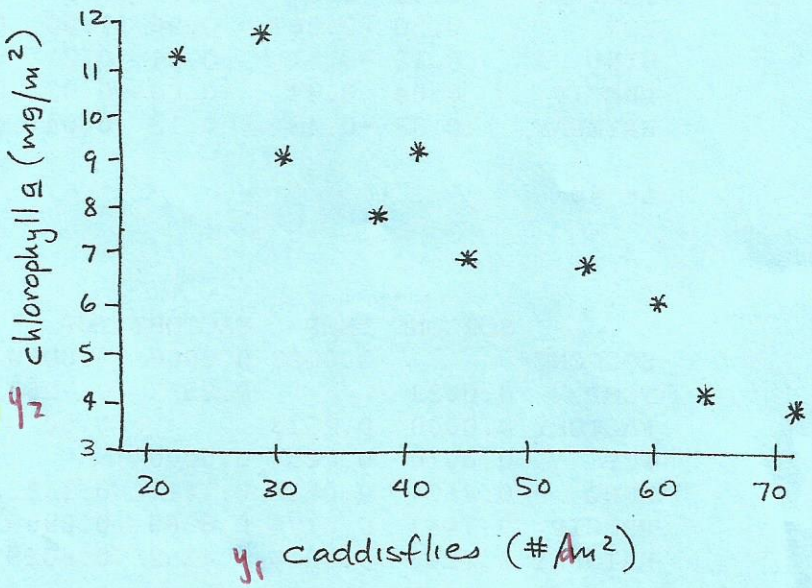
- How likely is it that a particular value of r occurred by chance alone when y_1 and y_2 are unrelated? ($H_0: \rho = 0$)
- I.N.E. assumptions apply
  - Independence – each y_{1i}, y_{2i} pair is random
  - Variables have bivariate, normal distr. (blue sheet)
  - Variance of two variables are equal
  - Relationship is Linear

CORRELATION ANALYSIS

4/10/2014

Data: The number of grazing caddisflies and the abundance of attached algae (as chlorophyll a) was determined from ten rocks within a stream pool. Is there a relationship between caddisflies and algae?

Random Rock	Y_1 caddis (#/dm²)	Y_2 Chl a (mg/m²)
1	41	9.2
2	72	3.7
3	31	9.0
4	29	11.8
5	64	4.3
6	44	6.9
7	22	11.2
8	38	7.4
9	61	6.0
10	55	6.7
$\bar{y}$	45.7	7.62
	$s_{y_1}^2$	$s_{y_2}^2$
$\Sigma(y_i-\bar{y})^2$	2488.1	65.12



$\Sigma_i(y_{1i}-\bar{y}_1)(y_{2i}-\bar{y}_2) = S_{y_1y_2} = -376.64$ ($s_{y_1y_2}$)

$$r = \frac{-376.64}{\sqrt{2488.1}\sqrt{65.12}} = -0.936$$

(s_{y_1}, s_{y_2})

pretty strong negative correlation

---

test $H_0: \rho = 0$

$S_r^2 = \frac{1-r^2}{n-2}$

$r^2 = 0.936^2$

* tells us if correlation is significant

$$t = \frac{|-0.936 - 0|}{\sqrt{\frac{1-0.876}{10-2}}} = 7.55$$

(S_r; # bivariate points)

$t = \frac{r-0}{S_r}$

$t_{0.05,\nu=8} = 2.306$ And significant!

$\therefore$ reject H_0 @ $p < 0.001$

REGRESSION: assuming # caddis flies are fixed

$$\hat{y} = a + bx$$

$$\text{chl a} = 14.54 - 0.151(\text{caddis})$$

$$r = b\frac{S_{y_1}}{S_{y_2}} = -0.151\frac{\sqrt{2488.1}}{\sqrt{65.12}} = -0.93$$

this always works for SLR

Pearson's correlation matrix (R output)

	SO2CONC	TEMP	FACTORY	POP	WIND	PRECIP	RAINDAY
SO2CONC	1.00	-0.47	0.65	0.50	0.12	0.04	0.38
TEMP	-0.47	1.00	-0.19	-0.06	-0.31	0.37	-0.43
FACTORY	0.65	-0.19	1.00	0.96	0.24	-0.03	0.13
POP	0.50	-0.06	0.96	1.00	0.21	-0.02	0.04
WIND	0.12	-0.31	0.24	0.21	1.00	0.01	0.15
PRECIP	0.04	0.37	-0.03	-0.02	0.01	1.00	0.51
RAINDAY	0.38	-0.43	0.13	0.04	0.15	0.51	1.00

n= 40

P

	SO2CONC	TEMP	FACTORY	POP	WIND	PRECIP	RAINDAY
SO2CONC		0.0023	0.0000	0.0010	0.4604	0.7847	0.0165
TEMP	0.0023		0.2523	0.7257	0.0532	0.0176	0.0056
FACTORY	0.0000	0.2523		0.0000	0.1411	0.8666	0.4252
POP	0.0010	0.7257	0.0000		0.1882	0.8889	0.8029
WIND	0.4604	0.0532	0.1411	0.1882		0.9712	0.3714
PRECIP	0.7847	0.0176	0.8666	0.8889	0.9712		0.0009
RAINDAY	0.0165	0.0056	0.4252	0.8029	0.3714	0.0009	

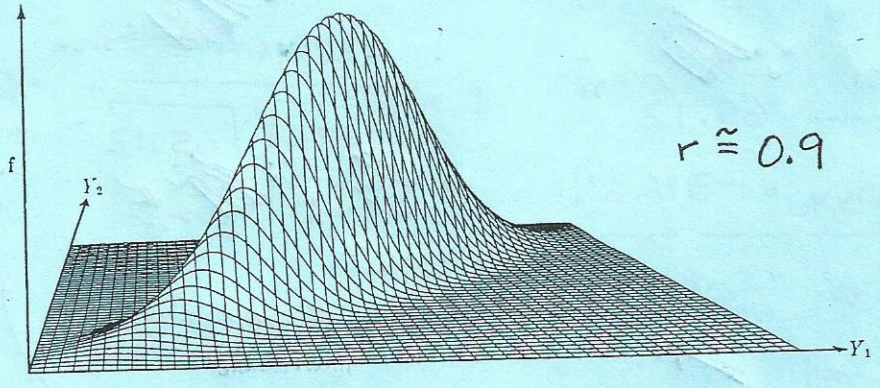


FIGURE 12.2
Bivariate normal frequency distribution. The parametric correlation ρ between variables Y_1 and Y_2 equals 0.9. The bell-shaped mound of Figure 12.1 has become elongated.

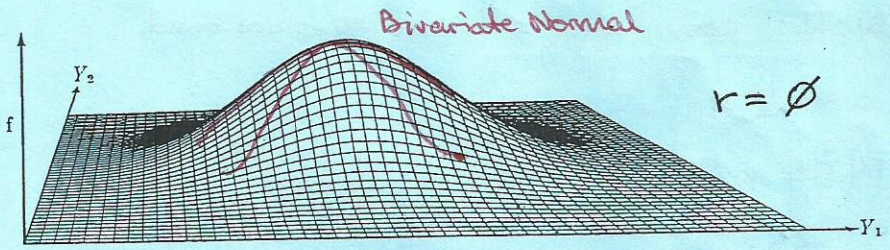


FIGURE 12.1
Bivariate normal frequency distribution. The parametric correlation ρ between variables Y_1 and Y_2 equals zero. The frequency distribution may be visualized as a bell-shaped mound.

4/10/2012

1. estimate of variance in r

$$S_r^2 = \frac{1-r^2}{n-2}$$ estimate of variable means (2 of them)

$$S_r = \sqrt{\frac{1-r^2}{n-2}}$$ standard error on r

2. Conduct significance test

$$t = \frac{r-0}{S_r} \sim t_{\alpha, n-2(df)}$$

3. $t = \frac{r-0}{S_r}$ (corr.) $= \frac{b-0}{S_b}$ (regr.) correlation related to regression mathematically

REMEMBER:

Regression is meant to be a predictive tool

Correlation is meant to show linear association b/w 2 variables

EXAM 3 Tuesday

- New index card, both sides
- 1/3 def., 2/3 short answers

1. Transformations
   a. Log, $\sqrt{}$, arcsin$\sqrt{}$, reciprocal, square (exponent)
2. Linear Regression
   a. Unreplicated Simple Linear Regression
      i. $H_0: \beta = 0$, $H_0: \alpha = 0$, R^2
   b. Replicated Linear Regression (SLR)
      i. $H_0: \beta = 0$, $H_0: \alpha = 0$, H_0: linearity, R^2 (test whether linear model fits)
   c. Differences b/w slopes, intercepts
      i. 2 slopes: t-tests; >2: ANCOVA (didn't do math for this)
   d. Multiple Regression (two or more predictors)
      i. All regressions, backstep, forstep, stepwise
3. Non-Linear Regression
   a. Polynomial (actually linear w/ multiple slopes)
   b. Exponential (linearize w/ log)
   c. Michaelis-Menton eq. (enzyme kinetics, linearize w/ reciprocal)
   d. Logistic Equation
4. Analysing Residuals
   a. Predictors, predictions, Time (NPP) → detect violations of assumptions
5. Correlation (Y_1, Y_2)
   a. Pearson's r; $H_0: \rho = 0$ (correlation = 0)

Final Exam

4/17/14

Non-parametric Statistics

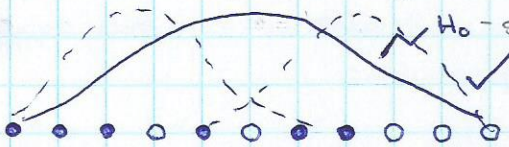
- Used when the assumptions of ANOVA cannot be met, even after transformations
- Also called "distribution free" tests - not based on normal distributions
  - ↳ But somewhat affected by egregious deviation from normal
- Most parametric tests have non-parametric equivalent
- We will only discuss a few of these tests

Non-parametric Tests we will cover

Parametric Test	Equivalent Test →	Non-parametric Test
Independent t-test		Mann-Whitney U-test
Paired t-test		Wilcoxon signed-rank test
One-way ANOVA		Kruskall-Wallis Test
Two-way ANOVA		Friedman Test
T-test of 2-distributions		Kolmogorov Smirnov test

Mann-Whitney U-test for 2 independent groups

- Substitutes for independent t-test
- Equivalent to Wilcoxon 2 sample test
- Null hypothesis H_0: two groups do not differ
  - Not looking specifically at means
- Alternative hypothesis: two groups differ
- Combined samples are ranked and the ranks of the smaller sample are summed
  - ↳ Pink sheet example ● = drug ○ = placebo



  - ↳ Ranking on pink sheet (smallest → largest)
- n_1 = smaller of two groups
- n_2 = larger of two groups
- Compute U and U' and choose the larger

Formula:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum_{i=1}^{n_1} R_{1i}$$ — sum of ranks of smaller sample

$$U' = n_1 n_2 - U'$$

- Compare U_{n_1,n_2} probabilities on table
- Ties receive mid-rank

4/17/14

- For pink sheet example — drug v. placebo

$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum_{i=1}^{n} R_{1j}$ → $\Sigma R = 4.5+6+9+10+11$

$= (5)(6) + \frac{(5)(5+1)}{2} - 40.5$

$= 30 + 15 - 40.5 = 4.5$

$U' = n_1 n_2 - U = (5)(6) - 4.5 = \underline{25.5 \text{ (larger than } U)}$

↳ Look at table $U_{0.05, n_1, n_2}$ → $U_{0.05, 5, 6} = 27$

$U < U_{crit}$ — fail to reject H_0 $0.10 < p < 0.05$

Wilcoxon Signed Rank Test (T)

Substitutes for paired t-test

Considers both the sign and magnitude of the difference b/w paired observations in the ranking

↳ See pink sheet again → twins problem — sum ranks for positives, negatives

$n = 12$ comparisons $\alpha = 0.05$ — ties recieve midrank like M-W

Wilcoxon $T^+ = \Sigma R_{positives} = 1+2+3+4+7+8+9+10+11+12 = 67$

$T^- = \Sigma R_{negatives} = 6+5 = 11$ (don't use signs here)

Take smaller of two T's (in this case $T^- = 11$)

— Use as test statistic #

Test statistic T must be smaller than critical value to be considered significant

$T_{crit} = T_{\alpha, n} = T_{0.05, 12} = 13$

$T_{crit} > T$ Reject H_0 $0.05 > p > 0.2$

If one of paired measurements $d_i = 0$ → drop these measurements and decrease value of n

- Notice we are not using any degrees of freedom — this is because we are not estimating any parameters

WILCOXON SIGNED-RANK TEST

4/17/2014

Example: Birth weights of identical ♂ twins (H_0: no difference)

First born	5.06	5.56	4.25	6.19	6.31	5.87	4.12	4.50	5.50	5.75	6.56	4.88
Second born	4.88	5.25	4.06	6.12	6.44	5.31	4.06	4.62	5.12	5.64	4.50	4.83
d_i	0.18	0.31	0.19	0.07	−0.13	0.56	0.06	−0.12	0.38	0.11	2.06	0.05
$\lvert d_i \rvert$	0.18	0.31	0.19	0.07	0.13	0.56	0.06	0.12	↓	↓	↓	↓
rank	order + rank 2 + put signs											
signed rank	0.05	0.06	0.07	0.11	0.12	0.13	0.18	0.19	0.31	0.38	0.56	2.06
	1	2	3	4	−5	−6	7	8	9	10	11	12

Negative sign added back on

Note: test statistic must be less than critical value to be significant

TABLE B.12 Critical Values of the Wilcoxon *T* Distribution

$\alpha(2) = 0.50$	0.20	0.10	0.05	0.02	0.01	0.005	0.001
$\alpha(1) = 0.25$	0.10	0.05	0.025	0.01	0.005	0.0025	0.0005
n							
4	2	0					
5	4	2	0				
6	6	3	2	0			
7	9	5	3	2	0		
8	12	8	5	3	1	0	
9	16	10	8	5	3	1	0
10	20	14	10	8	5	3	1
11	24	17	13	10	7	5	3	0
12	29	21	17	13	9	7	5	1
13	35	26	21	17	12	9	7	2
14	40	31	25	21	15	12	9	4
15	47	36	30	25	19	15	12	6
16	54	42	35	29	23	19	15	8
17	61	48	41	34	27	23	19	11
18	69	55	47	40	32	27	23	14
19	77	62	53	46	37	32	27	18
20	86	69	60	52	43	37	32	21
21	95	77	67	58	49	42	37	25
22	104	86	75	65	55	48	42	30
23	114	94	83	73	62	54	48	35
24	125	104	91	81	69	61	54	40
25	136	113	100	89	76	68	60	45
26	148	124	110	98	84	75	67	51
27	160	134	119	107	92	83	74	57
28	172	145	130	116	101	91	82	64
29	185	157	140	126	110	100	90	71
30	198	169	151	137	120	109	98	78

Mann-Whitney U-test for Two Independent Samples

Example: Systolic blood pressure for ♀ hypertensive patients. One group received a placebo; the other group received a new blood pressure drug. Data are blood pressures after 1 week of therapy. H_0: drug has no effect.

smaller group — smaller n

Group 1 (placebo): 165, 181, 159, 178, 192 ($n_1 = 5$)

Group 2 (drug): 151, 168, 140, 172, 138, 159 ($n_2 = 6$)

larger group — larger n

Source	2	2	2	1	2	1	2	2	1	1	1
Value	138	140	151	159	159	165	168	172	178	181	192
Rank	1	2	3	4.5		6	7	8	9	10	11

When 2 samples tie
↳ intermediate ranking

Note: test statistic must exceed critical value to be significant

TABLE B.11 (cont.) Critical Values of the Mann-Whitney *U* Distribution

	α(2):	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
	α(1):	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
n_1	n_2								
4	32	91	98	104	110	114	117	120	122
	33	94	101	107	113	117	120	124	126
	34	96	104	110	116	120	124	127	130
	35	99	107	113	120	124	127	131	133
	36	102	110	116	123	127	131	135	137
	37	105	113	119	126	131	134	138	141
	38	107	116	122	130	134	138	142	144
	39	110	118	125	133	137	141	145	148
4	40	113	121	129	136	141	145	149	152
n_1 (5)	5	20	21	23	24	25	--	--	--
n_2 (6)	6	23	25	27	28	29	30	--	--
	7	24	29	30	32	34	35	--	--
	8	30	32	34	36	38	39	40	--
	9	33	36	38	40	42	43	44	45
	10	37	39	42	44	46	47	49	50
	11	40	43	46	48	50	52	53	54
	12	43	47	49	52	54	56	58	59
	13	47	50	53	56	58	60	62	63
	14	50	54	57	60	63	64	67	68
	15	53	57	61	64	67	69	71	72
	16	57	61	65	68	71	73	75	77
	17	60	65	68	72	75	77	80	81
	18	63	68	72	76	79	81	84	86
	19	67	72	76	80	83	86	88	90
	20	70	75	80	84	87	90	93	95
	21	73	79	83	88	91	94	97	99
	22	77	82	87	92	96	98	102	104
	23	80	86	91	96	100	103	106	108
	24	84	90	95	100	104	107	110	113
	25	87	93	98	104	108	111	115	117
	26	90	97	102	108	112	115	119	121
	27	94	100	106	112	119	120	123	126
	28	97	104	110	116	120	124	128	130
	29	100	107	113	120	124	128	132	135
	30	104	111	117	124	128	132	136	139

4/22/2014

Advantages of Non-Parametrics

1. Not dependent on normality or equal variance and other assumptions of ANOVA
   - ↳ however, to some degree there cannot be too much deviation
2. Can organize, analyze ordinal scale data, such as ranks
3. Ease of calculation
4. Protection against outliers or gross measurement errors that ruin ANOVA

Disadvantage of Non-Parametrics

1. Loss of information because observations are replaced (usually by ranks)
2. Not as powerful as parametric statistics ($1-\beta$ declines)
   - ↳ Note: Assumption of independence of observations remains in force!
   - Also groups are assumed to have same distribution, even if not "normal"

Kruskall-Wallis Test for k independent groups

- Substitutes for one-way ANOVA
- Replace observations w/ ranks (low to high) without regard to original group
- Ties receive midrank
  - ↳ Green sheet
- For each group, sum the ranks ($= R_i$)
- Compute Kruskall-Wallis 'H' stat
- Compare to critical H in table — $H_{crit} = H_{\alpha, n_1, n_2, n_3}$ the group sample sizes are used to find crit. value

$$H = \left[\frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} \right] - 3(N+1)$$

(12 and 3 are constants)

↳ The H table we have is not perfect → doesn't have all sets of possible sample sizes

Kruskall Wallis Notes

- Can be used for 2 groups in place of U-test w/ similar results
- Non-parametric contrasts can also be performed

Friedman Test for 2-way classifications

- In place of multiway ANOVAs, randomized blocks ANOVA
- Developed by Milton Friedman – economist
  - ↳ yellow sheet
- Similar to Kruskall-Wallis, except observations are ranked for each block or level of the other factor (block) separately
- Compute Friedman's χ^2_r statistic
- Compare to critical χ^2_r in table

$$\chi^2_r = \left[\frac{12}{b \cdot k(k+1)} \sum_{i=1}^{k} R_i^2 \right] - 3 \cdot b(k+1)$$

k = groups b = blocks

4/22/14

Friedman requirements and notes

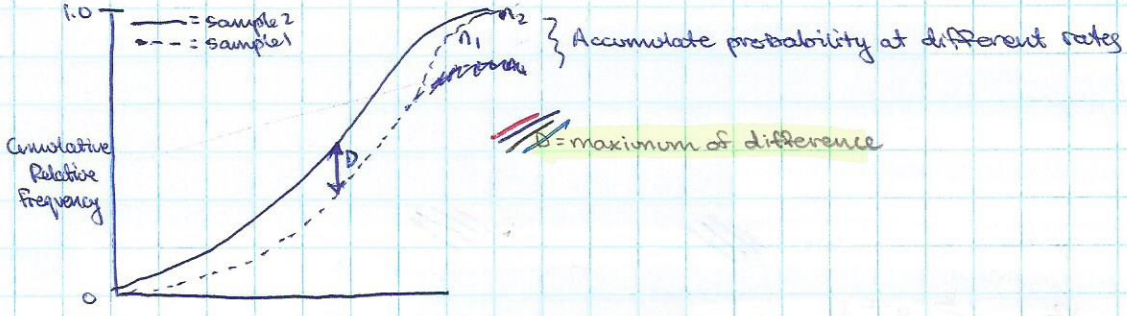
- Easily extended to examine 2 factors by generating Ri's for both columns and row's
  - Ex. for yellow sheet homework treat groves as second factor and rank w/in methods (1-7)
- Must be balanced
- Factor interaction cannot be computed
- Can accomodate >1 observation per cell (replicates)
- Non-parametric contrasts also exists

Kolmogorov-Smirnov Test for 2 distributions

Some non-parametric tests don't use ranks but instead test differences between two distributions

H_0: two distributions are the same

The K-S test compares two cumulative frequency distributions



Performing K-S test

Arrange data in alternating ascending order

Compute the relative cumulative frequencies for each sample

Compute difference d_i between the two frequencies

Identify maximum difference D and multiply by the 2 sample sizes ($D \cdot n_1 \cdot n_2$)

Compare to critical K-S for 2 sample sizes

Features of K-S test

Alternative to U-test but lower power

Some consider better than U-test because dispersion and location considered

K-S also can be used for testing an observed distribution to an expected distribution (normal, Poisson) so serves as a non-parametric "goodness of fit" test

4/22/14

KRUSKAL-WALLIS TEST for k Independent Groups

Example: Pollen load (g) of honeybees foraging at different distances from the hive. $k = 3$ distances.

H_0: no difference in pollen load

H_a: difference in pollen load w/ distance

Distance:	200 m		100 m		10 m		
	x_{1j}	rank	x_{2j}	rank	x_{3j}	rank	
x_{ij}	0.34	1	1.438		2.67	15	
	1.05	3	0.652		2.06	12	
	1.18	5	2.514		3.51	17	
	1.83	11	1.499		2.87	16	
	1.20	6	1.114		3.72	18	
	1.56	10			4.68	19	
	1.32	7			2.41	13	$N = 19$
n_i	7		5		7		
R_i	43		37		110		
R_i^2	1849		1369		12100		

Computing the K-W statistic H:

$$H = \left[\frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i}\right] - 3(N+1) = \left[\frac{12}{19(19+1)}\left[\frac{1849}{7} + \frac{1369}{5} + \frac{12100}{7}\right]\right] - 3(19+1)$$

$H_{crit} < H$

critical $H_{.05,7,5,7} \cong 5.819$ $\therefore$ Reject H_0 with $pH = 11.57$

α ; 7, 5, 7 = group sample sizes

p-value < 0.001

TABLE B.12 CRITICAL VALUES OF THE KRUSKAL-WALLIS H DISTRIBUTION

n_1	n_2	n_3	α = 0.10	0.05	0.02	0.01	0.005	0.002	0.001
2	2	2	4.571						
3	2	1	4.286						
3	2	2	4.500	4.714					
3	3	1	4.571	5.143					
3	3	2	4.556	5.361	6.250				
3	3	3	4.622	5.600	6.489	(7.200)	7.200		
4	2	1	4.500						
4	2	2	4.458	5.333	6.000				
4	3	1	4.056	5.208					
4	3	2	4.511	5.444	6.144	6.444	7.000		
4	3	3	4.709	5.791	6.564	6.745	7.318	8.018	
4	4	1	4.167	4.967	(6.667)	6.667			
4	4	2	4.555	5.455	6.600	7.036	7.282	7.855	
4	4	3	4.545	5.598	6.712	7.144	7.598	8.227	8.909
4	4	4	4.654	5.692	6.962	7.654	8.000	8.654	9.269
5	2	1	4.200	5.000					
5	2	2	4.373	5.150	6.000	6.533			
5	3	1	4.018	4.960	6.044				
5	3	2	4.651	5.251	6.124	6.909	7.182		
5	3	3	4.533	5.648	6.533	7.079	7.636	8.048	8.727
5	4	1	3.987	4.985	6.431	6.955	7.364		
5	4	2	4.541	5.273	6.505	7.205	7.573	8.114	8.591
5	4	3	4.549	5.656	6.676	7.445	7.927	8.481	8.795
5	4	4	4.619	5.657	6.953	7.760	8.189	8.868	9.168
5	5	1	4.109	5.127	6.145	7.309	8.182		
5	5	2	4.623	5.338	6.446	7.338	8.131	6.446	7.338
5	5	3	4.545	5.705	6.866	7.578	8.316	8.809	9.521
5	5	4	4.523	5.666	7.000	7.823	8.523	9.163	9.606
5	5	5	4.940	5.780	7.220	8.000	8.780	9.620	9.920
5	1	1	-----						
5	2	1	4.200	4.822					
6	2	2	4.545	5.345	6.182	6.982			
5	3	1	3.909	4.855	6.236				
5	3	2	4.682	5.348	6.227	6.970	7.515	8.182	
6	3	3	4.538	5.615	6.590	7.410	7.872	8.628	9.346
6	4	1	4.038	4.947	6.174	7.106	7.614		
6	4	2	4.494	5.340	6.571	7.340	7.846	8.494	8.827
5	4	3	4.604	5.610	6.725	7.500	8.033	8.918	9.170
5	4	4	4.595	5.681	6.900	7.795	8.381	9.167	9.861
6	5	1	4.128	4.990	6.138	7.182	8.077	8.515	
6	5	2	4.596	5.338	6.585	7.376	8.196	8.967	9.189
5	5	3	4.535	5.602	6.829	7.590	8.314	9.150	9.669
5	5	4	4.522	5.661	7.018	7.936	8.643	9.458	9.960
6	5	5	4.547	5.729	7.110	8.028	8.859	9.771	10.271
5	5	1	4.000	4.945	6.286	7.121	8.165	9.077	9.692
6	6	2	4.438	5.410	6.667	7.467	8.210	9.219	9.752
6	6	3	4.558	5.625	6.900	7.725	8.458	9.458	10.150
5	6	4	4.548	5.724	7.107	8.000	8.754	9.662	10.342
6	6	5	4.542	5.765	7.152	8.124	8.987	9.948	10.524
6	6	6	4.643	5.801	7.240	8.222	9.170	10.187	10.889
7	7	7	4.594	5.819	7.332	8.378	9.373	10.516	11.310
8	8	8	4.595	5.805	7.355	8.465	9.495	10.805	11.705
2	2	1	1 -----						
2	2	2	1 5.357	5.679					
2	2	2	2 5.667	6.167	(6.667)	6.667			

FRIEDMAN TEST for 2-WAY CLASSIFICATION

4/22/14

· Example: Pesticide residue on leaves is a hazard for fieldworkers. Pesticide was applied to 7 different groves of oranges using 3 different application procedures. Use a randomized blocks design to determine if residues ($\mu g/cm^2$ of leaf) differ among the different application methods. $k = 3$ methods, $b = 7$ blocks

H_0 = no diff among methods

	Method A		Method B		Method C			
Grove	X_{1j}	Rank	X_{2j}	Rank	X_{3j}	Rank	$\bar{X}_j$	$R_i = 13 \Rightarrow 169$
block 1	281	2	274	1	313	3	289.3	
2	143	1	196	3	171	2	170.0	$= 5 \Rightarrow 25$
3	473	1	492	2	624	3	529.7	$= 21 \Rightarrow 441$
4	122	1	141	2	180	3	147.7	$= 4 \Rightarrow 16$
5	251	1	315	2	336	3	300.7	$= 14 \Rightarrow 196$
6	386	1	413	3	394	2	397.7	$= 18 \Rightarrow 324$
7	173	1	268	2	299	3	246.7	$= 9 \Rightarrow 81$
$\bar{X}_i$	261.3		299.6		331.0		296.4	
R_i	8		15		19			
R_i^2	64		225		361			

Computing Friedman's χ_r^2:

$$\chi_r^2 = \left[\frac{12}{b \cdot k(k+1)} \sum_{i=1}^{k} R_i^2\right] - 3 \cdot b(k+1) = \left[\frac{12}{7 \cdot 3(3+1)} \left[(64 + 225 + 361)\right]\right] - 3(7)(3+1)$$

$$\chi_r^2 = 8.86$$

critical $\chi^2_{r\,.05,3,7} = 7.143$ $\therefore$ Reject H_0 with $p < 0.01$

(k, b)

w/ block as first factor

$$\chi^2_r = \left[\frac{12}{3 \cdot 7(7+1)} \Sigma\left[169 + 25 + 441 + 16 + 196 + 324 + 81\right]\right] - 3(3)(7+1) = 17.42$$

crit. $\chi^2_{r\,0.05,7,3} = \sim 9$ Reject H_0

★ If you had multiple replicates, you would rank each replicate individually

TABLE B.14 Critical Values of the Friedman χ_r^2 Distribution

k (n)	b (M)*	α = 0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
3	2	3.000	4.000							
3	3	2.667	4.667	(6.000)	6.000					
3	4	2.000	4.500	6.000	6.500	(8.000)	(8.000)	8.000		
3	5	2.800	3.600	5.200	6.400	(8.400)	8.400	(10.000)	(10.000)	10.000
3	6	2.330	4.000	5.330	7.000	8.330	9.000	(10.330)	10.330	12.000
3	7	2.000	3.714	5.429	7.143	8.000	8.857	10.286	11.143	12.286
3	8	2.250	4.000	5.250	6.250	7.750	9.000	9.750	12.000	12.250
3	9	2.000	3.556	5.556	6.222	8.000	9.556	10.667	11.556	12.667
3	10	1.800	3.800	5.000	6.200	7.800	9.600	10.400	12.200	12.600
3	11	4.636	3.818	4.909	6.545	7.818	9.455	10.364	11.636	13.273
3	12	1.500	3.500	5.167	6.167	8.000	9.500	10.167	12.167	12.500
3	13	1.846	3.846	4.769	6.000	8.000	9.385	10.308	11.538	12.923
3	14	1.714	3.571	5.143	6.143	8.143	9.000	10.429	12.000	13.286
3	15	1.733	3.600	4.933	6.400	8.133	8.933	10.000	12.133	12.933
4	2	3.600	5.400	(6.000)	6.000					
4	3	3.400	5.400	6.600	7.400	8.200	(9.000)	(9.000)	9.000	
4	4	3.000	4.800	6.300	7.800	8.400	9.600	(10.200)	10.200	11.100
4	5	3.000	5.160	6.360	7.800	9.240	9.960	10.920	11.640	12.600
4	6	3.000	4.800	6.400	7.600	9.400	10.200	11.400	12.200	12.800
4	7	2.829	4.886	6.429	7.800	9.343	10.371	11.400	12.771	13.800
4	8	2.550	4.800	6.300	7.650	9.450	10.350	11.850	12.900	13.800

k (n)	b (M)*	α = 0.10	0.05	0.025	0.01	0.005	0.001
4	9	6.467	7.800	9.133	10.867	12.067	14.467
4	10	6.360	7.800	9.120	10.800	12.000	14.640
4	11	6.382	7.909	9.327	11.073	12.273	14.891
4	12	6.400	7.900	9.200	11.100	12.300	15.000
4	13	6.415	7.985	7.369	11.123	12.323	15.277
4	14	6.343	7.886	9.343	11.143	12.514	15.257
4	15	6.440	8.040	9.400	11.240	12.520	15.400
5	2	7.200	7.600	8.000	8.000		
5	3	7.467	8.533	9.600	10.133	10.667	11.467
5	4	7.600	8.800	9.800	11.200	12.000	13.200
5	5	7.680	8.960	10.240	11.680	12.480	14.400
5	6	7.733	9.067	10.400	11.867	13.067	15.200
5	7	7.771	9.143	10.514	12.114	13.257	15.657
5	8	7.800	9.300	10.600	12.300	13.500	16.000
5	9	7.733	9.244	10.667	12.444	13.689	16.356
5	10	7.760	9.280	10.720	12.480	13.840	16.480
6	2	8.286	9.143	9.429	9.714	10.000	
6	3	8.714	9.857	10.810	11.762	12.524	13.286
6	4	9.000	10.286	11.429	12.714	13.571	15.286
6	5	9.000	10.486	11.743	13.229	14.257	16.429
6	6	9.048	10.571	12.000	13.619	14.762	17.048
6	7	9.122	10.674	12.061	13.857	15.000	17.612
6	8	9.143	10.714	12.214	14.000	15.286	18.000
6	9	9.127	10.778	12.302	14.143	15.476	18.270
6	10	9.143	10.800	12.343	14.299	15.600	18.514

KOLMOGOROV-SMIRNOV 2-SAMPLE TEST

4/22/14

Example: Mouthpart length for 2 samples of chiggers, in um.
Do they come from the same or different populations?

Ho: Same population

Ho: two frequency distributions are the same

Sample A X_{1i}	Sample B X_{2i}	$\frac{F_1}{n_1}$	$\frac{F_2}{n_2}$	$d = \left\lvert \frac{F_1}{n_1} - \frac{F_2}{n_2} \right\rvert$
—	100	0	.100	.100
104	—	.062	"	.038
—	105	"	.200	.138
—	107	"	.300	.238
—	107	"	.400	.338
—	108	"	.500	.438
109	—	.125	"	.375
—	111	"	.600	.475
112	—	.188	"	.412
114	—	.250	"	.350
116	116	.312	.700	.388
118	—	.374	"	.326
118	—	.438	"	.262
119	—	.500	"	.200
—	120	"	.800	.300
121	121	.562	.900	.338
123	123	.625	1.00	.375
125	—	.688	"	.312
126	—	.750	"	.250
126	—	.812	"	.188
128	—	.874	"	.126
128	—	.936	"	.064
128	—	1.00	"	.000
n_i 16	10			

data

Alternate, in order numerically ascending

Cumulative Relative Frequencies/Probabilities

Now we look for diff

$= |0.1 - 0.062|$

$= |0.2 - 0.062|$

$= |0.3 - 0.062|$ etc...

D = maximum difference

Same

6.2% for each

each measurment represents 16% of total probability

Test Statistic = $D \cdot n_1 \cdot n_2 = (0.475)(16)(10) = 76$

Critical value from Table XIII is 84 for $\alpha = 0.05$

∴ Fail to reject H_0

TABLE XIII

Critical values of the two-sample Kolmogorov-Smirnov statistic.

n_2

n_1	α	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
2	.05	-	-	-	-	-	-	16	18	20	22	24	26	26	28	30	32	34	36	38	38	40	42	44	46
	.025	-	-	-	-	-	-	-	-	-	-	24	26	28	30	32	34	36	38	40	40	42	44	46	48
	.01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	38	40	42	44	46	48	50
3	.05	-	-	-	15	18	21	21	24	27	30	30	33	36	36	39	42	45	45	48	51	51	54	57	60
	.025	-	-	-	-	18	21	24	27	30	30	33	36	39	39	42	45	48	51	51	54	57	60	60	63
	.01	-	-	-	-	-	-	-	27	30	33	36	39	42	42	45	48	51	54	57	57	60	63	66	69
4	.05	-	-	16	20	20	24	28	28	30	33	36	39	42	44	48	48	50	53	60	59	62	64	68	68
	.025	-	-	-	20	24	28	28	32	36	36	40	44	44	45	52	52	54	57	64	63	66	69	72	75
	.01	-	-	-	-	24	28	32	36	36	40	44	48	48	52	56	60	60	64	68	72	72	76	80	84
5	.05	-	15	20	25	24	28	30	35	40	39	43	45	46	55	54	55	60	61	65	69	70	72	76	80
	.025	-	-	20	25	30	30	32	36	40	44	45	47	51	55	59	60	65	66	75	74	78	80	81	90
	.01	-	-	-	25	30	35	35	40	45	45	50	52	56	60	64	68	70	71	80	80	83	87	90	95
6	.05	-	18	20	24	30	30	34	39	40	43	48	52	54	57	60	62	72	70	72	75	78	80	90	88
	.025	-	18	24	30	36	35	36	42	44	48	54	54	58	63	64	67	78	76	78	81	86	86	96	96
	.01	-	-	24	30	36	36	40	45	48	54	60	60	64	69	72	73	84	83	88	90	92	97	102	107
7	.05	-	21	24	28	30	42	40	42	46	48	53	56	63	62	64	68	72	76	79	91	84	89	92	97
	.025	-	21	28	30	35	42	41	45	49	52	56	58	70	68	73	77	80	84	86	98	96	98	102	105
	.01	-	-	28	35	36	42	48	49	53	59	60	65	77	75	77	84	87	91	93	105	103	108	112	115
8	.05	16	21	28	30	34	40	48	46	48	53	60	62	64	67	80	77	80	82	88	89	94	98	104	104
	.025	-	24	28	32	36	41	48	48	54	58	64	65	70	74	80	80	86	90	96	97	102	106	112	112
	.01	-	-	32	35	40	48	56	55	60	64	68	72	76	81	88	88	94	98	104	107	112	115	128	125
9	.05	18	24	28	35	39	42	46	54	53	59	63	65	70	75	78	82	90	89	93	99	101	106	111	114
	.025	-	27	32	36	42	45	48	63	60	63	69	72	76	81	85	90	99	98	100	108	110	115	120	123
	.01	-	27	36	40	45	49	55	63	63	70	75	78	84	90	94	99	108	107	111	117	122	126	132	135
10	.05	20	27	30	40	40	46	48	53	70	60	66	70	74	80	84	89	92	94	110	105	108	114	118	125
	.025	-	30	36	40	44	49	54	60	70	68	72	77	82	90	90	96	100	103	120	116	118	124	128	135
	.01	-	30	36	45	48	53	60	63	80	77	80	84	90	100	100	106	108	113	130	126	130	137	140	150

n_2

n_1	α	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
11	.05	22	30	33	39	43	48	53	59	60	77	72	75	82	84	89	93	97	102	107	112	121	119	124	129
	.025	-	30	36	44	48	52	58	63	68	77	76	84	87	94	96	102	107	111	116	123	132	131	137	140
	.01	-	33	40	45	54	59	64	70	77	88	86	91	96	102	106	110	118	122	127	134	143	142	150	154
12	.05	24	30	36	43	48	53	60	63	66	72	84	81	86	93	96	100	108	108	116	120	124	125	144	138
	.025	24	33	40	45	54	56	64	69	72	76	96	84	94	99	104	108	120	120	124	129	134	137	156	150
	.01	-	36	44	50	60	60	68	75	80	86	96	95	104	108	116	119	126	130	140	141	148	149	168	165
13	.05	26	33	39	45	52	56	62	65	70	75	81	91	89	96	101	105	110	114	120	126	130	135	140	145
	.025	26	36	44	47	54	58	65	72	77	84	84	104	100	104	111	114	120	126	130	137	141	146	151	158
	.01	-	39	48	52	60	65	72	78	84	91	95	117	104	115	121	127	131	138	143	150	156	161	166	172
14	.05	26	36	42	46	54	63	64	70	74	82	86	89	112	98	106	111	116	121	126	140	138	142	146	150
	.025	28	39	44	51	58	70	70	76	82	87	94	100	112	110	116	122	126	133	138	147	148	154	160	166
	.01	-	42	48	56	64	77	76	84	90	96	104	104	126	123	126	134	140	148	152	161	164	170	176	182
15	.05	28	36	44	55	57	62	67	75	80	84	93	96	98	120	114	116	123	127	135	138	144	149	156	160
	.025	30	39	45	55	63	68	74	81	90	94	99	104	110	135	119	129	135	141	150	153	154	163	168	175
	.01	-	42	52	60	69	75	81	90	100	102	108	115	123	135	133	142	147	152	160	168	173	179	186	195
16	.05	30	39	48	54	60	64	80	78	84	89	96	101	106	114	128	124	128	133	140	145	150	157	168	167
	.025	32	42	52	59	64	73	80	85	90	96	104	111	116	119	144	136	140	145	156	157	164	169	184	181
	.01	-	45	56	64	72	77	88	94	100	106	116	121	126	133	160	143	154	160	168	173	180	187	200	199
17	.05	32	42	48	55	62	68	77	82	89	93	100	105	111	116	124	136	133	141	146	151	157	163	168	173
	.025	34	45	52	60	67	77	80	90	96	102	108	114	122	129	136	153	148	151	160	166	170	179	183	190
	.01	-	48	60	68	73	84	88	99	106	110	119	127	134	142	143	170	164	166	175	180	187	196	203	207
18	.05	34	45	50	60	72	72	80	90	92	97	108	110	116	123	128	133	162	142	152	159	164	170	180	180
	.025	36	48	54	65	78	80	86	99	100	107	120	120	126	135	140	148	162	159	166	174	178	184	198	196
	.01	-	51	60	70	84	87	94	108	108	118	126	131	140	147	154	164	180	176	182	189	196	204	216	216
19	.05	36	45	53	61	70	76	82	89	94	102	108	114	121	127	133	141	142	171	160	163	169	177	183	187
	.025	38	51	57	66	76	84	90	98	103	111	120	126	133	141	145	151	159	190	169	180	185	190	199	205
	.01	38	54	64	71	83	91	98	107	113	122	130	138	148	152	160	166	176	190	187	199	204	209	218	224
20	.05	38	48	60	65	72	79	88	93	110	107	116	120	126	135	140	146	152	160	180	173	176	184	192	200
	.025	40	51	64	75	78	86	96	100	120	116	124	130	138	150	156	160	166	169	200	180	192	199	208	215
	.01	40	57	68	80	88	93	104	111	130	127	140	143	152	160	168	175	182	187	220	199	212	219	228	235
21	.05	38	51	59	69	75	91	89	99	105	112	120	126	140	138	145	151	159	163	173	189	183	189	198	202
	.025	40	54	63	74	81	98	97	108	116	123	129	137	147	153	157	166	174	180	180	210	203	206	213	220
	.01	42	57	72	80	90	105	107	117	126	134	141	150	161	168	173	180	189	199	199	231	223	227	237	244
22	.05	40	51	62	70	78	84	94	101	108	121	124	130	138	144	150	157	164	169	176	183	198	194	204	209
	.025	42	57	66	78	86	96	102	110	118	132	134	141	148	154	164	170	178	185	192	203	220	214	222	228
	.01	44	60	72	83	92	103	112	122	130	143	148	156	164	173	180	187	196	204	212	223	242	237	242	250
23	.05	42	54	64	72	80	89	98	106	114	119	125	135	142	149	157	163	170	177	184	189	194	230	205	216
	.025	44	60	69	80	86	98	106	115	124	131	137	146	154	163	169	179	184	190	199	206	214	230	226	237
	.01	46	63	76	87	97	108	115	126	137	142	149	161	170	179	187	196	204	209	219	227	237	253	249	262
24	.05	44	57	68	76	90	92	104	111	118	124	144	140	146	156	168	168	180	183	192	198	204	205	240	225
	.025	46	60	72	81	96	102	112	120	128	137	156	151	160	168	184	183	198	199	208	213	222	226	264	238
	.01	48	66	80	90	102	112	128	132	140	150	168	166	176	186	200	203	216	218	228	237	242	249	288	262
25	.05	46	60	68	80	88	97	104	114	125	129	138	145	150	160	167	173	180	187	200	202	209	216	225	250
	.025	48	63	75	90	96	105	112	123	135	140	150	158	166	175	181	190	196	205	215	220	228	237	238	275
	.01	50	69	84	95	107	115	125	135	150	154	165	172	182	195	199	207	216	224	235	244	250	262	262	300

4/22/2014

Final Notes

- Don't use non-parametrics to "rescue" bad data or bad experiment
- Use non-parametrics when:
  - when you have collected ordinal scale data (ranks)
  - Can't meet ANOVA assumptions
  - Prefer use
  - Feeling crazy

4/24/2014

Sample Size Estimation

1. Objective: What do I want to measure and will I manipulate factors?
2. Experimental Design: How do I set up my experiment
3. Sampling Design: How should I sample in space and time?
4. * Replication: How many samples should I take?

How do I best determine sample size?

- There can be too many samples, too few, or just the right amount
- By trial and error, such as plotting mean and 95% CI for different sample sizes and then eyeballing the graph
- Use sample size statistics

To estimate sample size:

Little reminder:

~~Normal~~ x_i $\quad P\left[(\bar{x} - t_{\alpha,\nu} s_{\bar{x}}) < \mu < (\bar{x} + t_{\alpha,\nu} s_{\bar{x}})\right] = 1-\alpha$ (95% CI)

$\bar{x} \pm t_{\alpha,\nu} s_{\bar{x}}$ — precision of estimate of $\bar{x}$ (d)

↑ what we want to know

1. Decide what level of precision around a mean is desired
   - ↳ $\bar{x} \pm$ 20% of true value, ±10%, ±5%, etc.
   - Percent error around mean can be described as relative error D

$$\text{relative error } D = \frac{\text{absolute error } (d)}{\bar{x}} \rightarrow d = D \cdot \bar{x}$$

2. Use an equation to relate n to that precision level

$$d = t_{\alpha,\nu} s_{\bar{x}}$$

$$d = \frac{t \cdot s}{\sqrt{n}}$$

Absolute error

$$n = \frac{t^2 s^2}{d^2} = \left(\frac{ts}{d}\right)^2$$

Also... $d = D \cdot \bar{x}$

$$n = \left(\frac{t \cdot s}{D \cdot \bar{x}}\right)^2$$ Relative Error

↳ But... you actually need to know the mean and SD beforehand to use these equations

4/24/14

3. Estimate parameters for the equation
   a) previous sampling of system or similar system
   b) pilot study of target pop. to find $\bar{x}$ and SD measures
   c) literature values

4. Solve the equation

Ex. Estimate length of 1yr old perch in Lake Erie

Want perch length $\bar{x} \pm 25\%$ (D)

- From previous information $\bar{x} = 11.5$ cm $s = 3.8$ cm

$d = D \cdot \bar{x} = 0.25(11.5) = 2.9$ cm

$n = \frac{t^2 s^2}{d^2}$ $t_{0.05, \nu} =$

↳ What do we do for degrees of freedom?

↳ If you look at t-table at 0.05, 2 tails, the value of t converges to 2

↳ Good place to start!

$n = \frac{(2)^2(3.8)^2}{(2.9)^2} = 6.87 \rightarrow$ At least 7 fish need to be caught to achieve a mean of $\bar{x} \pm 25\%$

↳ Part b/c we estimated t, $t_{0.05, 7-1=6} = 2.447$

$n = \frac{(2.447)^2(3.8)^2}{(2.9)^2} = 10.28 \rightarrow$ 11 fish now

↳ $t_{0.05, \nu = 10-1=10} = 2.28$, closer to t

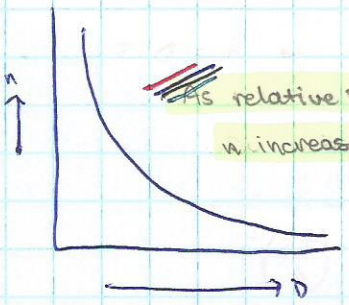
Additional iterations will zero in on good value of n

Ex. $\bar{x} \pm 10\%$

$d = 0.1(11.5) = 1.15$ cm

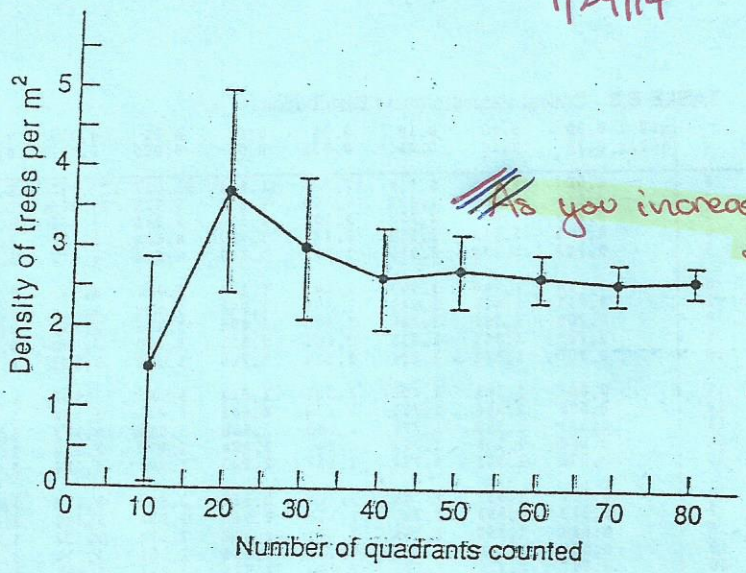
$n = \frac{(2)^2(3.8)^2}{(1.15)^2} = 43.67 \rightarrow 44$ fish

$t_{0.05, \nu = 44-1=43} = 2.017$ looks pretty close to our original value for t so 44 fish is probably close to sample size we need to estimate $\bar{x} \pm 10\%$



As relative D decreases, n increases quickly

4/24/14



As you increase sample size the means stabalize and the CI is small

You may want to use this as proxy for # of samples

Figure 5.3 An empirical approach to determining how large a sample to take. An ecology class counted red alder trees on an area undergoing secondary succession. After each 10 quadrats were counted, the mean and 95% confidence interval were plotted. Sampling was continued until the confidence interval was judged to be sufficiently small.

TABLE 5.2 COEFFICIENTS OF VARIATION OBSERVED IN A VARIETY OF POPULATION SAMPLING TECHNIQUES TO ESTIMATE POPULATION SIZE[a]

Group of organisms	Coefficient of variation
Aquatic organisms	
Plankton	0.70
Benthic organisms	
Surber sampler, counts	0.60
Surber sampler, biomass or volume	0.80
Grab samples or cores	0.40
Shellfish	0.40
Fish	0.50 to 2.00
Terrestrial organisms	
Roadside counts	0.80
Call counts	0.70
Transects (on foot)	0.50 to 2.00
Fecal pellet counts	1.00

Use CVs as estimates

[a] Average values compiled by Eberhardt (1978a).

TABLE B.3 Critical Values of the t Distribution

ν	$\alpha(2)$: 0.50 / $\alpha(1)$: 0.25	0.20 / 0.10	0.10 / 0.05	0.05 / 0.025	0.02 / 0.01	0.01 / 0.005	0.005 / 0.0025	0.002 / 0.001	0.001 / 0.0005
1	1.000	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
31	0.682	1.309	1.696	2.040	2.453	2.744	3.022	3.375	3.633
32	0.682	1.309	1.694	2.037	2.449	2.738	3.015	3.365	3.622
33	0.682	1.308	1.692	2.035	2.445	2.733	3.008	3.356	3.611
34	0.682	1.307	1.691	2.032	2.441	2.728	3.002	3.348	3.601
35	0.682	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591
36	0.681	1.306	1.688	2.028	2.434	2.719	2.990	3.333	3.582
37	0.681	1.305	1.687	2.026	2.431	2.715	2.985	3.326	3.574
38	0.681	1.304	1.686	2.024	2.429	2.712	2.980	3.319	3.566
39	0.681	1.304	1.685	2.023	2.426	2.708	2.976	3.313	3.558
40	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
41	0.681	1.303	1.683	2.020	2.421	2.701	2.967	3.301	3.544
42	0.680	1.302	1.682	2.018	2.418	2.698	2.963	3.296	3.538
43	0.680	1.302	1.681	2.017	2.416	2.695	2.959	3.291	3.532
44	0.680	1.301	1.680	2.015	2.414	2.692	2.956	3.286	3.526
45	0.680	1.301	1.679	2.014	2.412	2.690	2.952	3.281	3.520
46	0.680	1.300	1.679	2.013	2.410	2.687	2.949	3.277	3.515
47	0.680	1.300	1.678	2.012	2.408	2.685	2.946	3.273	3.510
48	0.680	1.299	1.677	2.011	2.407	2.682	2.943	3.269	3.505
49	0.680	1.299	1.677	2.010	2.405	2.680	2.940	3.265	3.500
50	0.679	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
1000	0.675	1.282	1.646	1.962	2.330	2.581	2.813	3.098	3.300
∞	0.6745	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.0902	3.2905

This table was prepared using Equations 26.7.3 and 26.7.4 of Zelen and Severo (1964), except for the values at infinity degrees of freedom, which are adapted from White (1970). Except for the values at infinity degrees of freedom, t was calculated to eight decimal places and then rounded to three decimal places.

Examples:

$$t_{0.05(2),13} = 2.160 \quad \text{and} \quad t_{0.01(1),19} = 2.539$$

4/24/2014

...But what if there is no previous info on population of interest?

$CV = \frac{s}{\bar{x}}$ useful metric for relative variation

$$n = \left(\frac{t_{\alpha,df}}{D\bar{x}}\right)^2 \approx \left(\frac{t}{D} CV\right)^2$$

↳ From blue sheet, use surbar samples $\bar{x} \pm 40\%$ $CV = 0.60$

$$n = \left[\frac{(2)}{(.4)}(0.6)\right]^2 = 9 \qquad t_{0.05,9-1=8} = 2.306$$

$$n = \left[\frac{2.306}{.4}(0.6)\right]^2 = 12 \qquad t_{0.05,12-1=11} = 2.173 \text{ etc...}$$

* At home try $\bar{x} \pm 20\%$

How do I deal w/ non-normal distributions?

- We were somewhat working on the assumption that distribution was normal
- Generally you can still use normal formulae because of central limit theorem that says means of sample approach normal distribution
- For highly skewed distributions, you may need to use a formula specific to the distribution (ex. negative binomial if skewed to the right)

$$n = \left[\left(\frac{2}{.2}\right)(0.6)\right]^2 = 36 \qquad t_{0.05,36-1=35} = 2.030$$

↳ need 36 replicates

4/29/2014

Pseudoreplication

- The use of inferential statistics to test for treatment effects when treatments are not replicated or replicates are not independent (Hurlbert 1984)
  - Can occur during "manipulative" or so called "mensurative" experiments
    - mensurative – observational

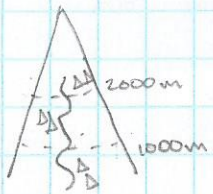
Experiments

Manipulative: an experiment where 2 or more treatments are imposed on experimental units

↳ Most of our examples this semester

Mensurative: an observational study where space or time serves as the comparison or "treatment"

↳ An example:

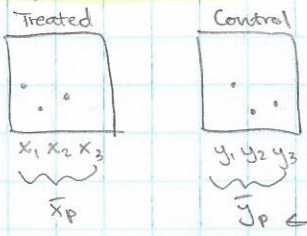


- Sampling tree types on mountain at different elevations, H_0: no difference b/w elevation

4/29/2014

Types of Pseudoreplication

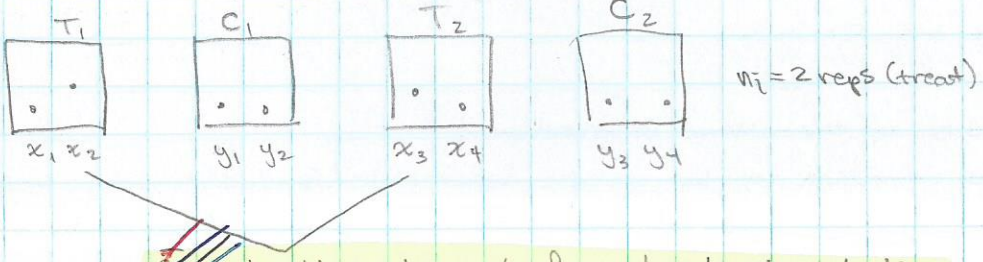
1. Simple: only a single replicate per treatment, but multiple samples taken from each treatment are taken as true replicates



← only subsamples of same treatment group

← better to take pooled mean

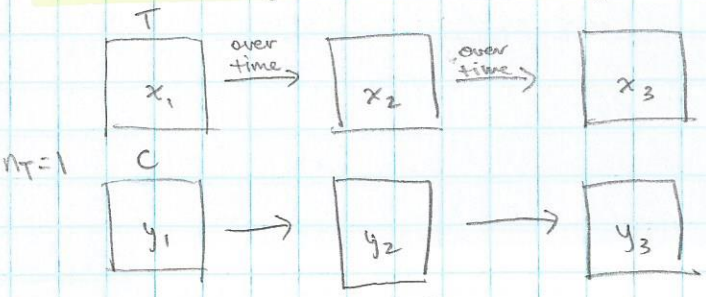
2. Sacrificial: treatments are truly replicated but then replicates are pooled (combined)



$n_i = 2$ reps (treat)

extracting subsamples from two treatments to get a less representative mean

3. Temporal: samples taken over time from the same experimental unit are (incorrectly) considered to be replicates

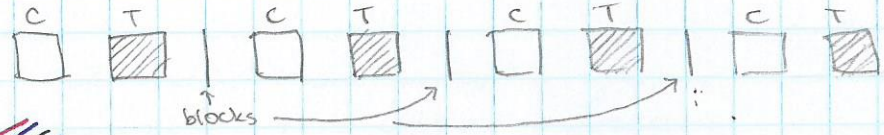


How to Randomize

1. Complete randomization — treatments assigned randomly to experimental units

- 8 plots, testing insecticide
  - ↳ If the second half were to be destroyed, we would need to start again b/c there would only be one T replicate

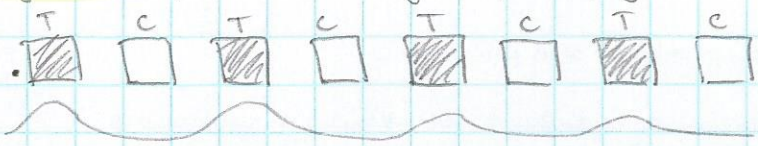
2. Interspersion (also known as stratified random or random block)



- Assign controls/treatments randomly w/in block
  - would be able to save experiment if second half destroyed b/c df = 1

4/29/14

2. Systematic placement – assign in pattern, system



- Problem is that there might be some oscillating condition that might influence outcome

  Ex. - treated plots at top of hills, controls at bottom

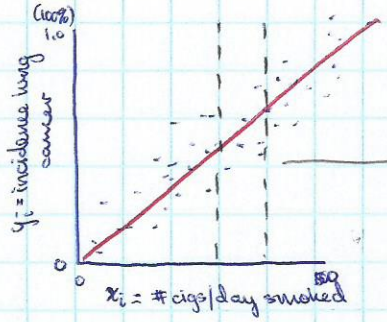
Happenstance Data

The inappropriate application of regression or correlation to data collected for a different purpose or without considering statistical assumptions

Some Types

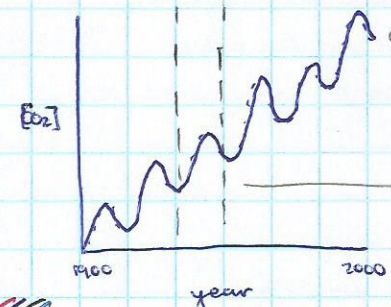
1. Limited range of variables

Thereby obscuring or hiding results



regression line obvious, high R^2

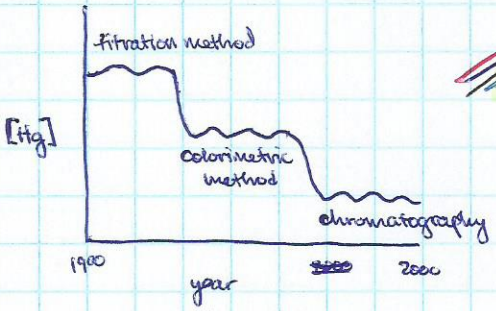
→ dishonest people (cig companies) may take subset and claim there is no association



Generally upward trend

→ subset misleading

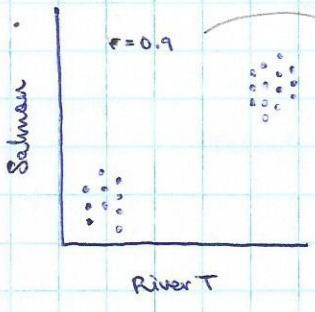
2. Inconsistent methods



- Not really difference due to decline, but due to detection limits of current measuring methods

4/29/2014

3. Extreme Data

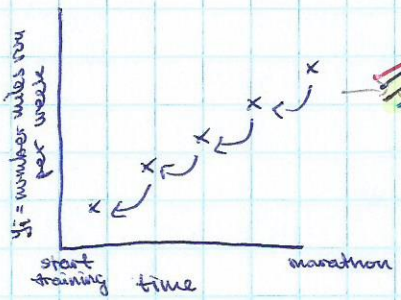


Could do correlation and get high r

→ but these are actually different salmon populations so not really drawing correct conclusion

4. Dependence on History or Serial Correlation

Measuring same unit over time

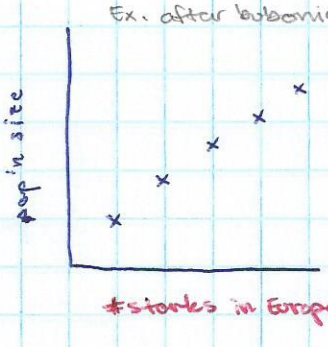


Every unit depends on the one before it

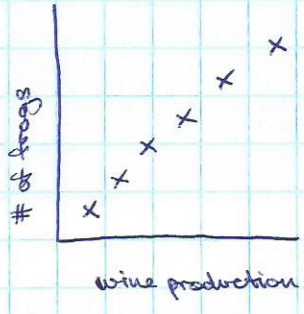
5. Nonsense correlation

relationship driven by hidden variable

Ex. after bubonic plague



- lurking variable was European prosperity
  - more people could afford to build chimney's which storks like to nest in



- lurking variable here is rainfall in the same year

THE END. :^

FINAL

FRIDAY 10:30 AM in 105 Jordan (same as class)

- Another index card; can bring all index cards
- Cumulative, but weighted towards most recent info covered (30%)

4/29/14

- FINAL EXAM MATERIAL

50% I. z distribution, t distribution, X^2 distribution } 1-2 questions

50% II ANOVA (all types) } 1-2 questions

50% III Regression (all forms), Correlation } 1-2 questions

30% IV Non-parametrics, Sample Size Statistics, Psuedoreplication / Happenstance } 2-3 Questions

20% Short answers (definitions/fill-ins)

RECENT MATERIAL - NON-PARAMETRICS

- Mann-Whitney U-test (ind. t-test)
- Wilcoxon signed Rank (paired t)
- Kruskal Wallis (one way ANOVA)
- Friedman Test (two way ANOVA)
- Kolmogorov-Smirnov (two distributions)

Sample size estimation

Psuedoreplication
- Simple
- Sacrificial
- Temporal

Happenstance
- Limited Range
- Inconsistent Methods
- Extreme Data
- Serial Correlation
- Nonsense Correlation