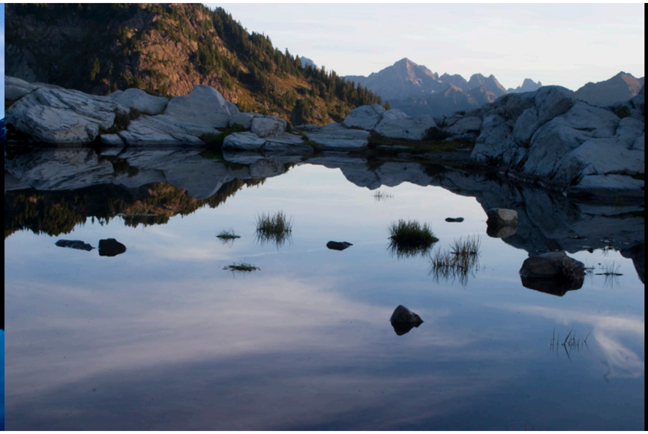
ANOVA, Correlation, Linear Regression, Quantile Regression


Jessica Lundquist, CEE 465/CEWA 565

October 10, 2019

Analysis of Variance (ANOVA) (like what we just did but for more than two different samples)

See Chapter 7 in the Helsel and Hirsch textbook

Can also check out Brandon Foltz on Youtube: https://www.youtube.com/watch?v=0Vj2V2qRU10

Note: We will go over One Factor ANOVA. There are many variations of ANOVA.

Note 2: This is cookbook like. Don't panic if you can't remember how many teaspoons of salt go in a given recipe. Just remember where you put the recipe and when you want to use it. Also consider when it's appropriate to use each test.

ANOVA (conceptually)

- Like a sequence of t-tests, but multiple t-tests end up compounding the type I error (end up with a higher level of error than you set up with alpha)
- Compares the overall variability among/ between means of different groups with the internal spread within each group

Illustration from: https://en.wikipedia.org/wiki/Analysis_of_variance

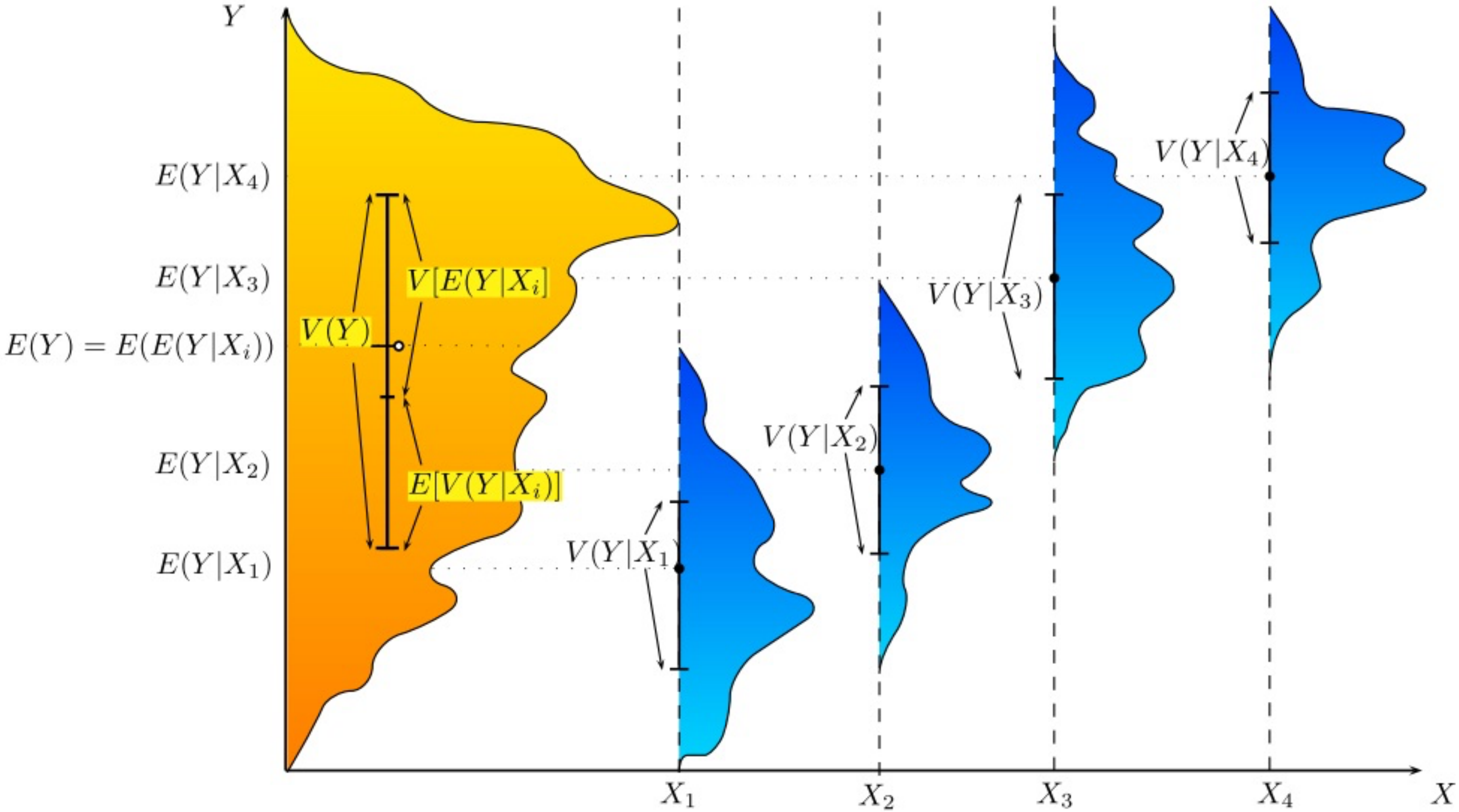


Figure 1: ANOVA : Fair fit

Illustration from: https://en.wikipedia.org/wiki/Analysis_of_variance

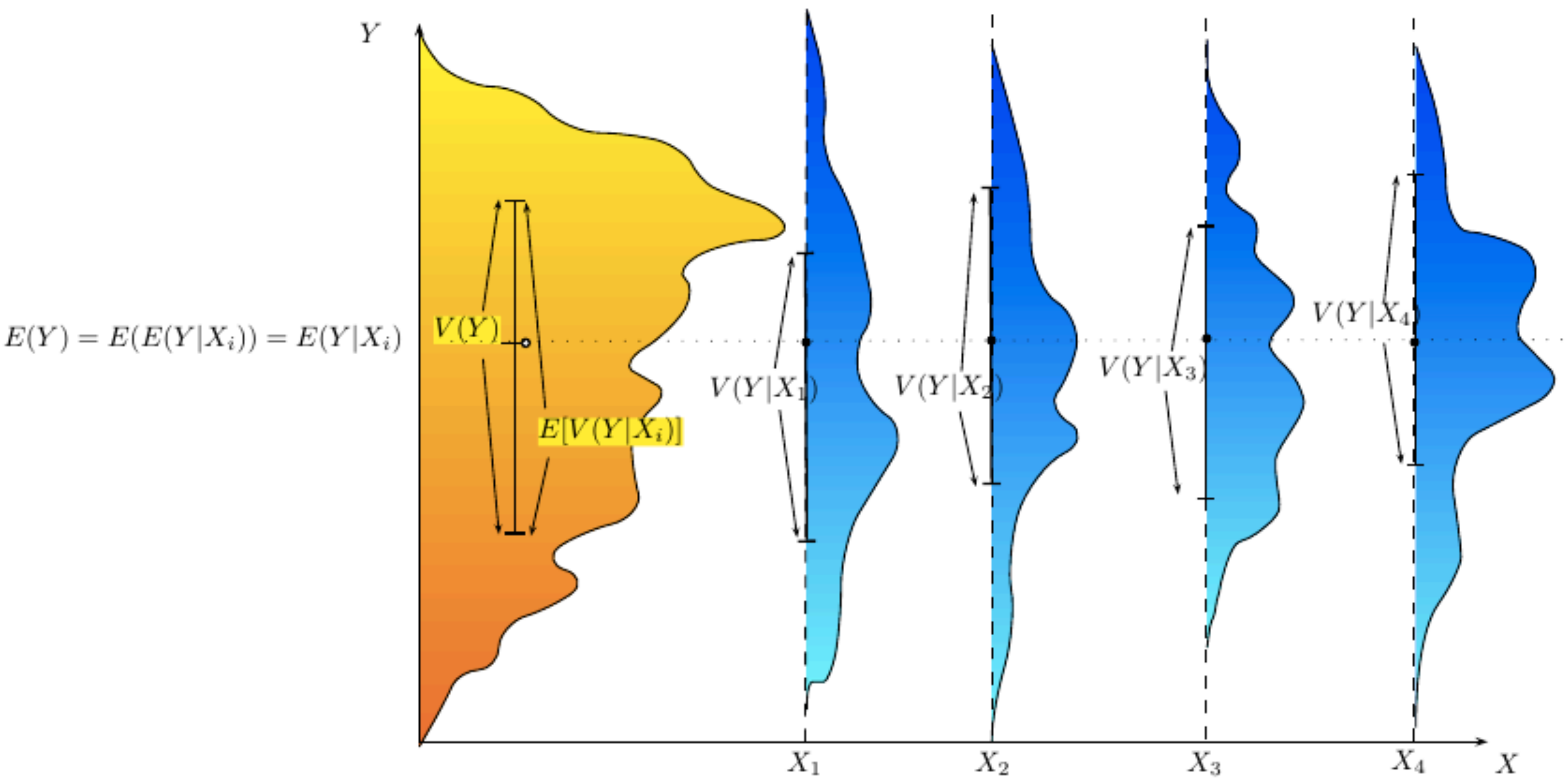


Figure 2: ANOVA : No fit

Illustration from: https://en.wikipedia.org/wiki/Analysis_of_variance

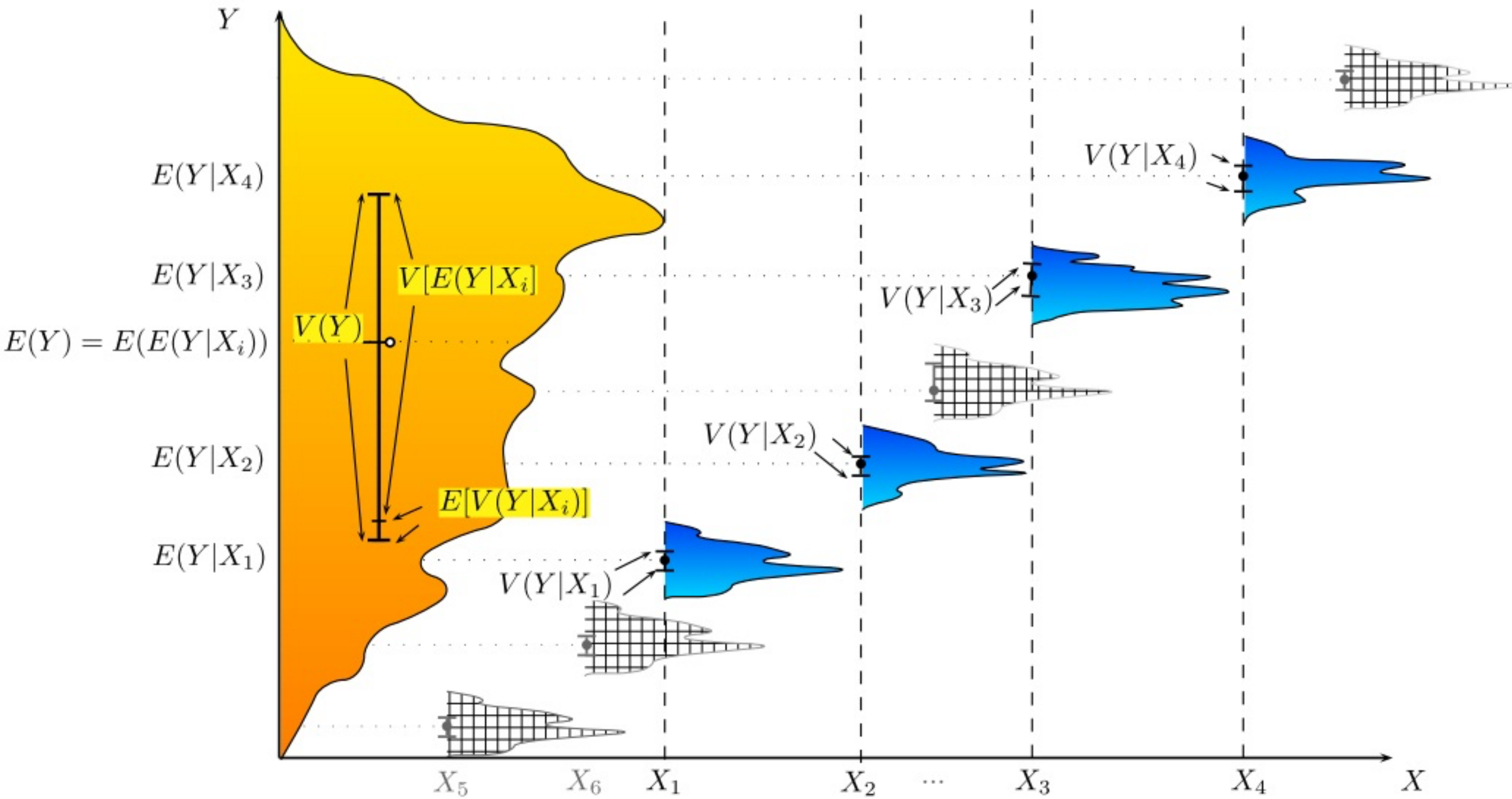
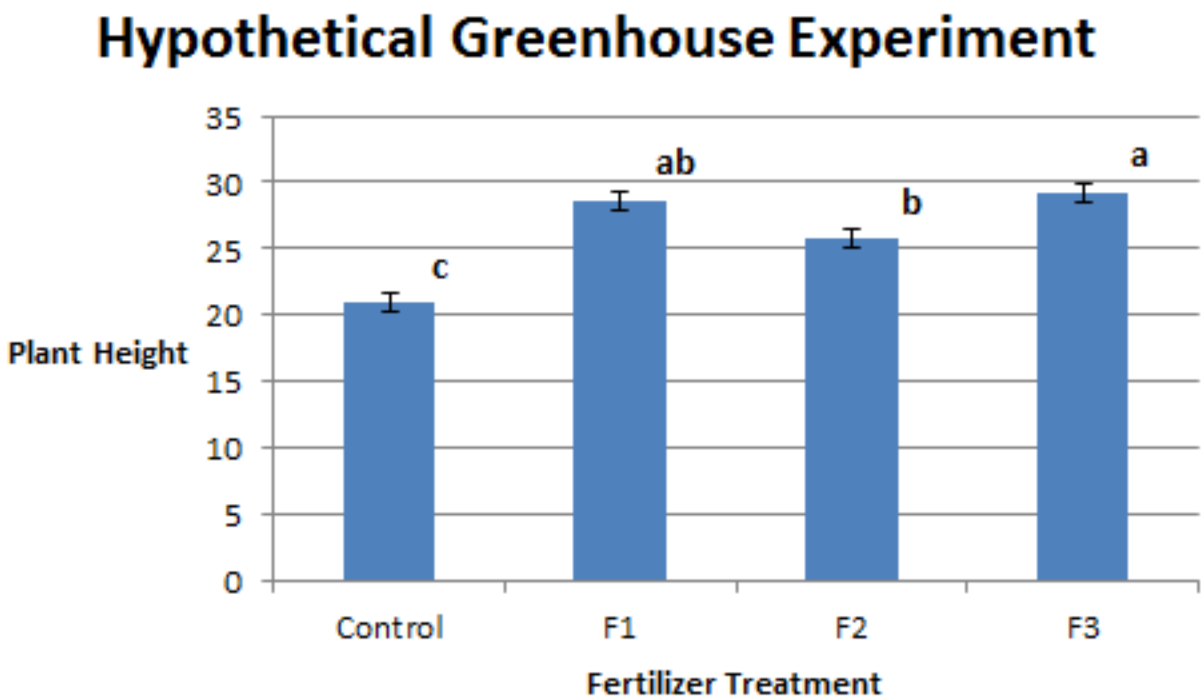


Figure 3: ANOVA : very good fit

We think plants will grow to different heights with different fertilizer treatment.

But we need a falsifiable hypothesis. What might this be?

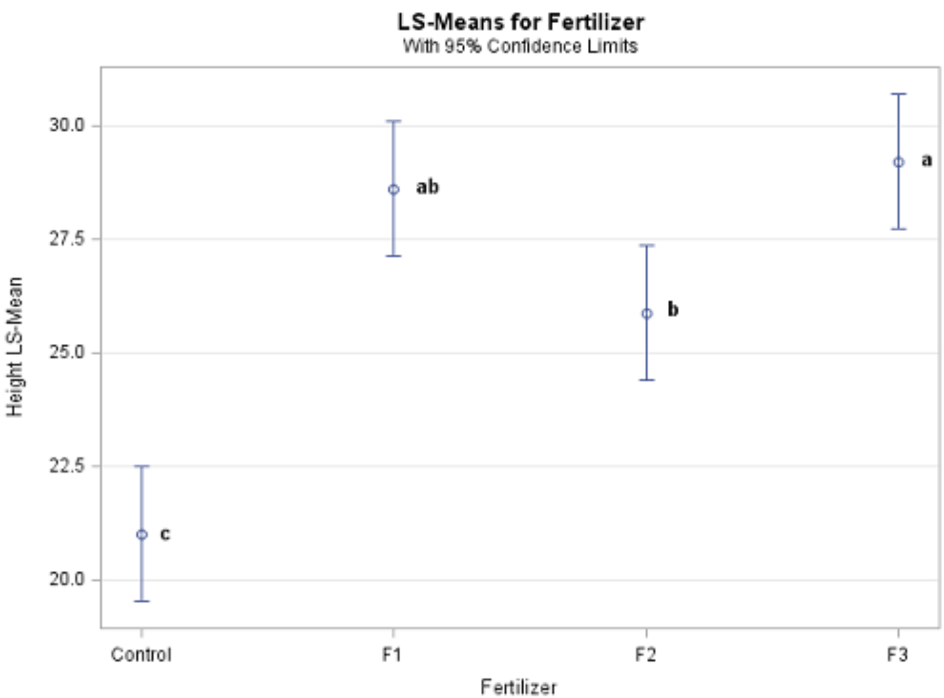


Graphic from https://onlinecourses.science.psu.edu/stat502/node/138

Null Hypothesis: All groups have the same central value (mean).

The test:

- Compare the mean values of each group with overall mean for the entire data set.
- Need to analyze the variance to tell if we can really tell the mean values apart
- The F-ratio looks at the variance between groups divided by the variance within groups



How much variability is due to the different groups vs. something else (error/within group variability)

Total sum of squares	=	Treatment sum of squares	+	Error sum of squares
(overall variation)	=	(group means – overall mean)	+	(variation within groups)

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 = \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2 + \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

If the total sum of squares is divided by N–1, where N is the total number of observations, it equals the variance of the y_{ij}'s. Thus ANOVA partitions the variance of the data into two parts, one measuring the signal and the other the noise. These parts are then compared to determine if the means are significantly different.

Where there are k different groups, with index j
And each group has n_j different samples, with index i

Key Assumptions for ANOVA

If ANOVA is performed on two groups, the F statistic which results will equal the square of the two-sample t-test statistic $F=t^2$, and will have the same p-value. It is not surprising, then, that the same assumptions apply to both tests:

1. All samples are random samples from their respective populations.
2. All samples are independent of one another.
3. Departures from the group mean $(y_{ij} - \bar{y}_j)$ are normally distributed for all j groups.
4. All groups have equal population variance σ^2 estimated for each group by s_j^2

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2}{n_j - 1}$$

See page 166 Helsel and Hirsh

Data from 3 Fertilizer Treatments (also in lab)

Fertilizer Treatment #

Height of plant (cm)

	j=1	j=2	j=3	j=4
	Control	**F1**	**F2**	**F3**
i=1	21	32	22.5	28
i=2	19.5	30.5	26	27.5
	22.5	25	28	31
	21.5	27.5	27	29.5
	20.5	28	26.5	30
i=6	21	28.6	25.2	29.2

Mean Square		Formula	Estimates:
Variance of y_{ij}	=	Total SS / N−1	Total variance of the data
MST	=	SST / k−1	Variance within groups + variance between groups.
MSE	=	SSE / N−k	Variance within groups.

N=24 (total number of observations)
K=4 (number of different groups)

ANOVA Table

Source	df	SS	MS	F	p-value
Treatment	(k−1)	SST	MST	MST/MSE	p
Error	(N−k)	SSE	MSE		
Total	N−1	Total SS			

See https://onlinecourses.science.psu.edu/stat502/node/137 for online lecture notes on ANOVA

Data from 3 Fertilizer Treatments (also in lab)

	j=1 Control	j=2 F1	j=3 F2	j=4 F3
i=1	21	32	22.5	28
i=2	19.5	30.5	26	27.5
	22.5	25	28	31
	21.5	27.5	27	29.5
	20.5	28	26.5	30
i=6	21	28.6	25.2	29.2

Mean Square		Formula	Estimates:
Variance of y_{ij}	=	Total SS / N−1	Total variance of the data
MST	=	SST / k−1	Variance within groups + variance between groups.
MSE	=	SSE / N−k	Variance within groups.

N=24
K=4

ANOVA

Source	df	SS	MS	F

See https://onlinecourses.science.psu.edu/stat502/node/137 for online lecture notes on ANOVA

Data from 3 Fertilizer Treatments (also in lab)

	j=1 Control	j=2 F1	j=3 F2	j=4 F3
i=1	21	32	22.5	28
i=2	19.5	30.5	26	27.5
	22.5	25	28	31
	21.5	27.5	27	29.5
	20.5	28	26.5	30
i=6	21	28.6	25.2	29.2

Mean Square		Formula	Estimates:
Variance of y_{ij}	=	Total SS / N−1	Total variance of the data
MST	=	SST / k−1	Variance within groups + variance between groups.
MSE	=	SSE / N−k	Variance within groups.

N=24
K=4

ANOVA

Source	df	SS	MS	F
Treatment	k-1=3			
Error	N-k=20			
Total	N-1=23			

See https://onlinecourses.science.psu.edu/stat502/node/137 for online lecture notes on ANOVA

Data from 3 Fertilizer Treatments (also in lab)

	j=1	j=2	j=3	j=4
	Control	**F1**	**F2**	**F3**
i=1	21	32	22.5	28
i=2	19.5	30.5	26	27.5
	22.5	25	28	31
	21.5	27.5	27	29.5
	20.5	28	26.5	30
i=6	21	28.6	25.2	29.2

Mean Square		Formula	Estimates:
Variance of y_{ij}	=	Total SS / N−1	Total variance of the data
MST	=	SST / k−1	Variance within groups + variance between groups.
MSE	=	SSE / N−k	Variance within groups.

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

$$SSTrt = 6 * (21.0 - 26.1667)^2 + 6 * (28.6 - 26.1667)^2 + 6 * (25.8667 - 26.1667)^2 + 6 * (29.2 - 26.1667)^2 = 251.44$$

ANOVA

Source	**df**	**SS**	**MS**	**F**
Treatment	k-1=3	251.44		
Error	N-k=20			
Total	N-1=23			

See https://onlinecourses.science.psu.edu/stat502/node/137 for online lecture notes on ANOVA

Data from 3 Fertilizer Treatments (also in lab)

	j=1	j=2	j=3	j=4
	Control	**F1**	**F2**	**F3**
i=1	21	32	22.5	28
i=2	19.5	30.5	26	27.5
	22.5	25	28	31
	21.5	27.5	27	29.5
	20.5	28	26.5	30
i=6	21	28.6	25.2	29.2

Mean Square		Formula	Estimates:
Variance of y_{ij}	=	Total SS / N−1	Total variance of the data
MST	=	SST / k−1	Variance within groups + variance between groups.
MSE	=	SSE / N−k	Variance within groups.

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

ANOVA

Source	**df**	**SS**	**MS**	**F**
Treatment	k-1=3	251.44		
Error	N-k=20	61.033		
Total	N-1=23			

See https://onlinecourses.science.psu.edu/stat502/node/137 for online lecture notes on ANOVA

Data from 3 Fertilizer Treatments (also in lab)

	j=1	j=2	j=3	j=4
	Control	**F1**	**F2**	**F3**
i=1	21	32	22.5	28
i=2	19.5	30.5	26	27.5
	22.5	25	28	31
	21.5	27.5	27	29.5
	20.5	28	26.5	30
i=6	21	28.6	25.2	29.2

Mean Square		Formula	Estimates:
Variance of y_{ij}	=	Total SS / N−1	Total variance of the data
MST	=	SST / k−1	Variance within groups + variance between groups.
MSE	=	SSE / N−k	Variance within groups.

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

Total sum of squares = **Treatment sum of squares** + **Error sum of squares**
(overall variation) = (group means − overall mean) + (variation within groups)

Recall: If you calculate two SS, you can get the third easily.

Source	df	SS	MS	F
Treatment	k-1=3	251.44		
Error	N-k=20	61.033		
Total	N-1=23	312.47		

See https://onlinecourses.science.psu.edu/stat502/node/137 for online lecture notes on ANOVA

Data from 3 Fertilizer Treatments (also in lab)

	j=1 Control	j=2 F1	j=3 F2	j=4 F3
i=1	21	32	22.5	28
i=2	19.5	30.5	26	27.5
	22.5	25	28	31
	21.5	27.5	27	29.5
	20.5	28	26.5	30
i=6	21	28.6	25.2	29.2

Mean Square		Formula	Estimates:
Variance of y_{ij}	=	Total SS / N−1	Total variance of the data
MST	=	SST / k−1	Variance within groups + variance between groups.
MSE	=	SSE / N−k	Variance within groups.

$$F = \frac{MS_{Trt}}{MS_{Error}} = \frac{83.813}{3.052} = 27.46$$

ANOVA

Source	df	SS	MS	F
Treatment	k-1=3	251.44	83.813	27.46
Error	N-k=20	61.033	3.052	
Total	N-1=23	312.47		

See https://onlinecourses.science.psu.edu/stat502/node/137 for online lecture notes on ANOVA

Look F up with a table or a software program. If it's p value is less than your rejection value, you reject the null.

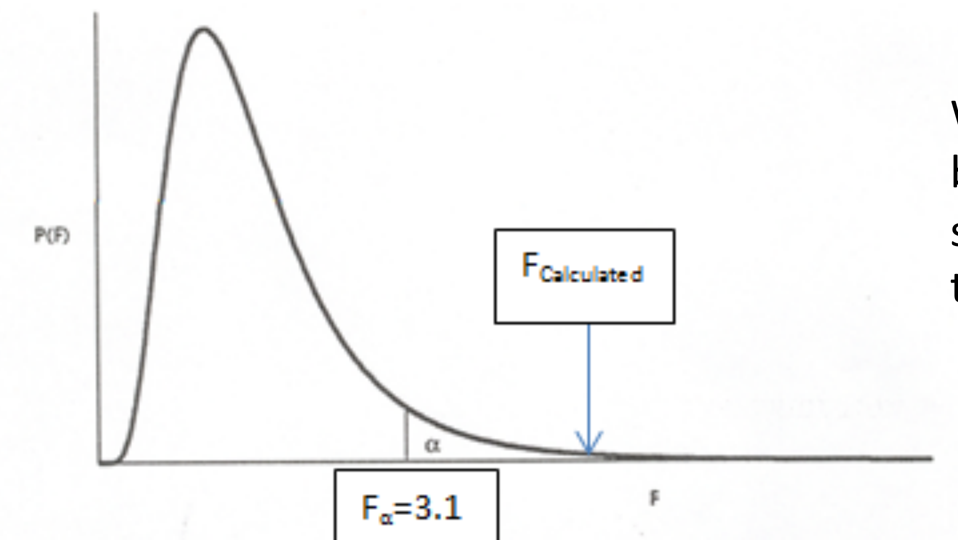


Figure K.1: The F distribution

We reject the null when the between treatment variance is significantly more than the within treatment variance.

See https://onlinecourses.science.psu.edu/stat502/node/137 for online lecture notes on ANOVA

One factor analysis of variance

Situation Several groups of data are to be compared, to determine if their means are significantly different. Each group is assumed to have a normal distribution around its mean. All groups have the same variance.

Computation The treatment mean square and error mean square are computed as their sum of squares divided by their degrees of freedom (df). When the treatment mean square is larger than the error mean square as measured by an F-test, the group means are significantly different.

$$\mathrm{MST} = \frac{\sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2}{k-1}$$ where k−1 = treatment degrees of freedom

$$\mathrm{MSE} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2}{N-k}$$ where N−k = error degrees of freedom

Tied data No alterations necessary.

Test Statistic The test statistic F:

F = MST / MSE

From Helsel and Hirsh

Decision Rule To reject H_0: the mean of every group is identical, versus

H_1: at least one mean differs .

Reject H_0 if $F \geq F_{1-\alpha,\, k-1,\, N-k}$ the 1−α quantile of an F distribution with k−1 and N−k degrees of freedom; otherwise do not reject H_0.

If you reject the null, how do you know which sample (or samples) differ(s) from the rest?

- Multiple Approaches Exist
- Check what your software is using and cite it appropriately
- Don't just use multiple t-tests because you can inflate your type I error
- Most common (recommended) is Tukey's test:

Two group means $\bar{y}_i$ and $\bar{y}_j$ can be considered different if

$$\left|\bar{y}_i - \bar{y}_j\right| > q_{(1-\alpha),\, k,\, N-k} \cdot \sqrt{MSE / n}$$

where

where	q	is the studentized range statistic from Neter, Wasserman and Kutner (1985),
	α	is the overall significance level,
	k	is the number of treatment group means compared,
	$N-k$	are the degrees of freedom for the MSE, and
	n	is the sample size per group.

Also online and in stats books as Tukey q values

See page 198 of H&H or https://onlinecourses.science.psu.edu/stat502/node/143

Also see: https://www.mathworks.com/help/stats/analysis-of-variance-and-covariance.html

See Ch 7 H&H for Non-parametric

- ANOVA is like a t-test between three or more groups of data, and as such is restricted to the same assumptions as the t-test
- The Kruskal-Wallis test is similar to the rank-sum test, but is extended to more than two groups and compares the medians.
- There are many more advanced versions of the ANOVA (e.g., Multi-factor and ANCOVA), which you may wish to read about, but we will not go into them in this class

Lab 3.1 teaches ANOVA

- Teaches you how to do the fertilizer problem we just did in python code
- Relevant to the first problem on homework 3

Correlation Analysis

Correlation Coefficient:

For two samples X and Y with sample size n , consider the quantity:

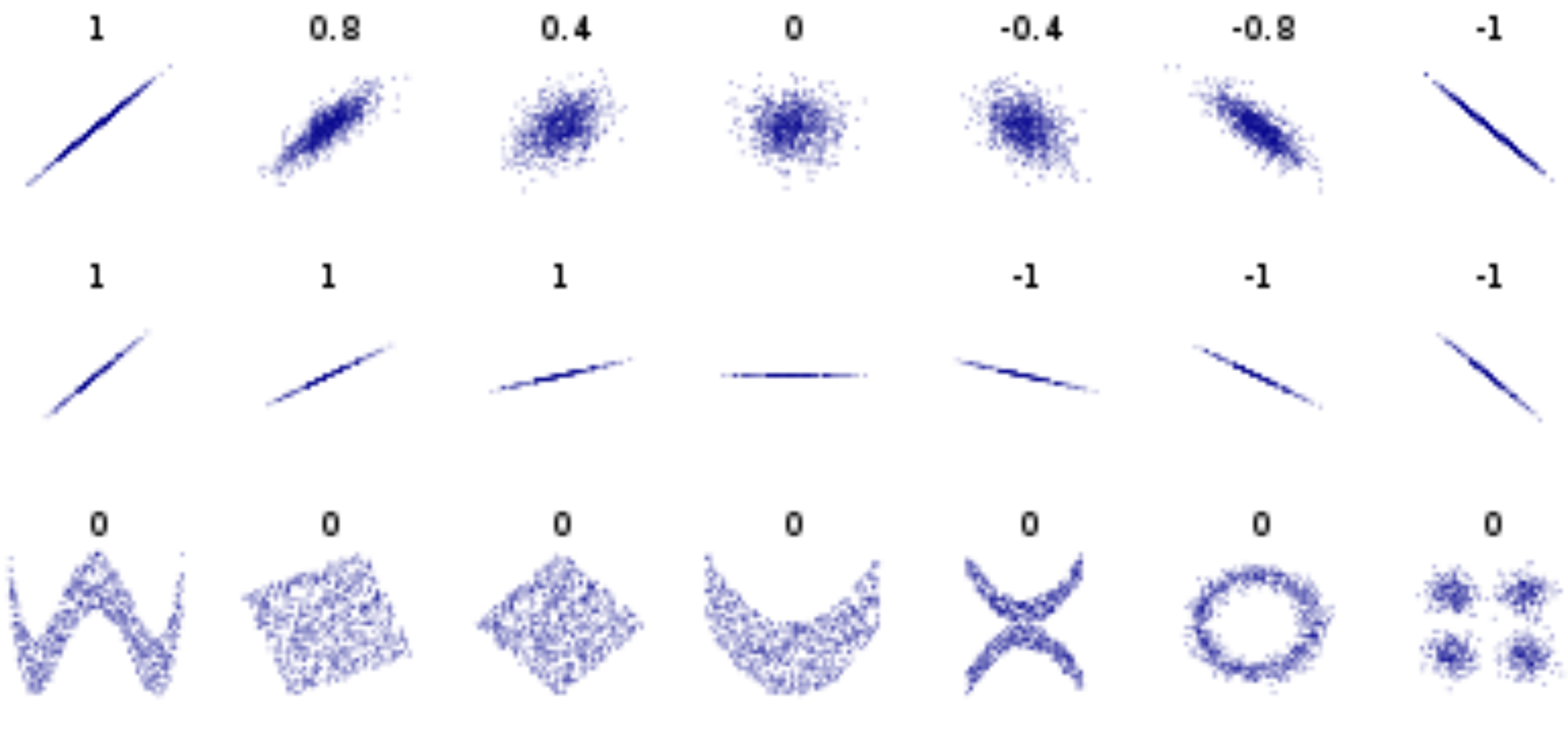
$$s_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

If the anomalies of X and Y are mostly in the same direction at the same time, this will be a large positive value. If the anomalies are mostly of opposite sign at the same time, this will be a large negative value. If there is essentially no pattern, this value will be small.

We can scale this quantity so it varies between -1.0 and 1.0 to define the correlation coefficient r:

$$r = \frac{s_{xy}}{\left(\sqrt{\sum (x_i - \bar{x})^2}\right)\left(\sqrt{\sum (y_i - \bar{y})^2}\right)}$$

Example Data Sets Plotted as (x,y) Pairs with Associated Correlation Coefficients



http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

Population vs. Sample Statistics

This is TRUE correlation, summed over all values

$$\rho = R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

So the sample statistic r from the previous slide is an estimator for the population statistic R above.

$$r = \frac{s_{xy}}{\left(\sqrt{\sum(x_i - \bar{x})^2}\right)\left(\sqrt{\sum(y_i - \bar{y})^2}\right)}$$

$$s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Use, for example, the CORREL function in Excel to calculate the sample correlation coefficient.

In matlab, use corrcoef (see the lab)

This is calculated for our dataset, sometimes also represented by R.

Some Important Characteristics of the Correlation Coefficient

1. Value of **r** does not depend on labeling of “x” and “y”.

2. The value of **r** is independent of units (and is also invariant for certain kinds of linear transformations).

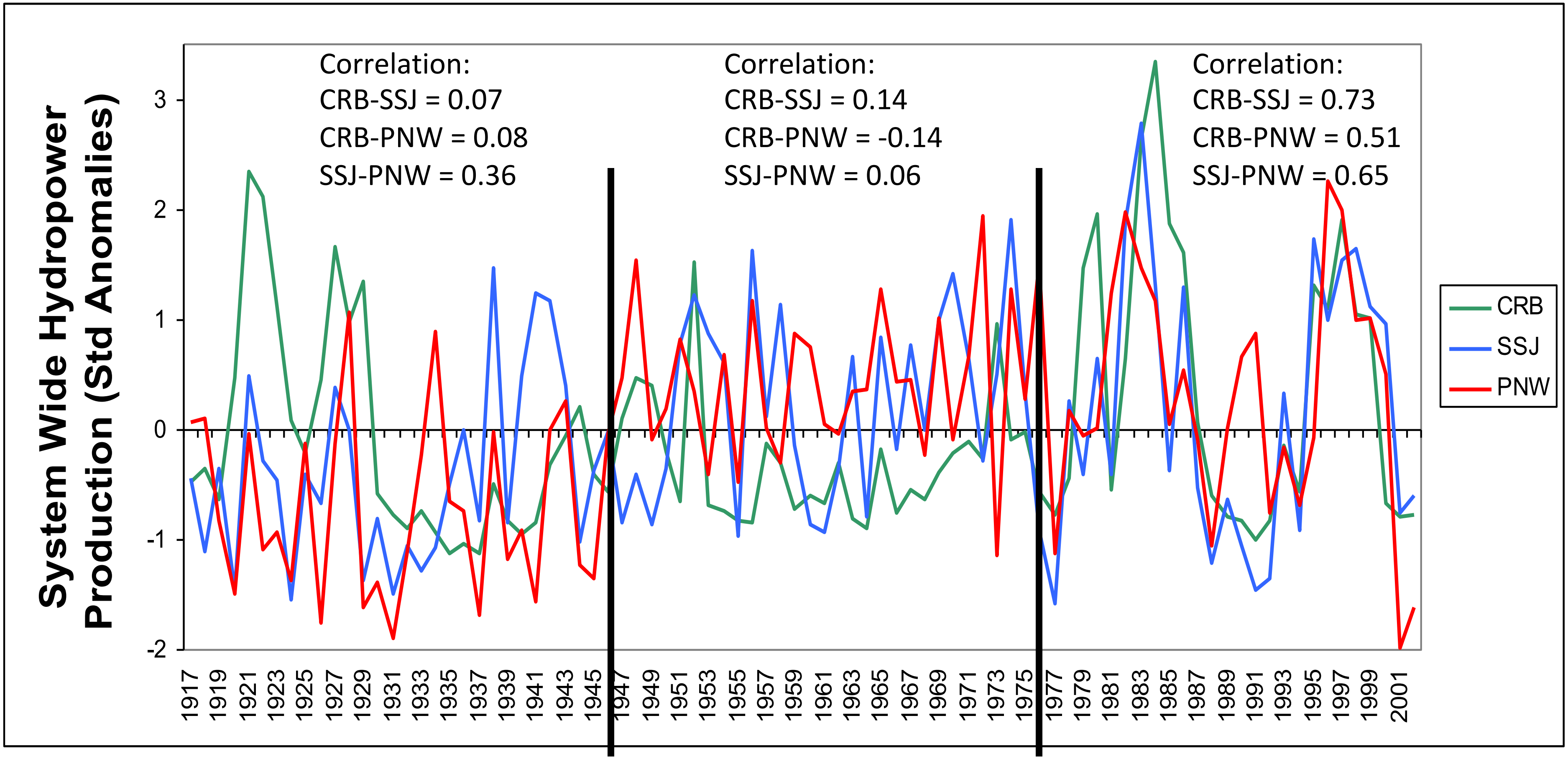
3. **r** lies in the interval [-1,1].

4. **r** = 1 if and only if the x,y pairs lie on a straight line with positive slope, and **r** = -1 if and only if the x,y pairs lie on a straight line with negative slope. If the slope is zero, then r = 0. (Note this a “problem” in that if the relationship is very weak, but strongly linear, the relationship is still reported as being very strong. Solution, plot the data!)

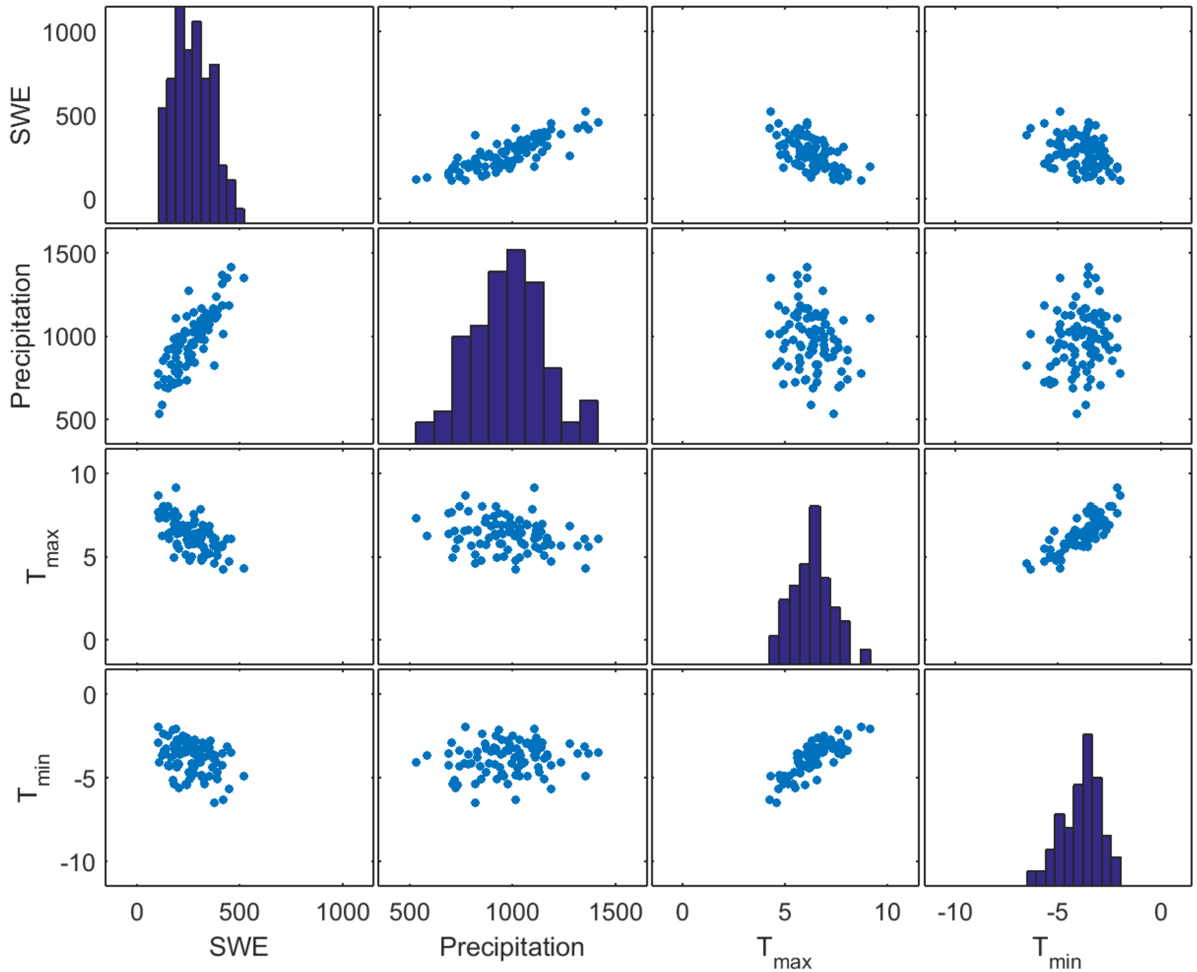
5. The square of the sample correlation coefficient is R^2, the coefficient of determination for a simple linear regression model between the two variables.

Example:

CRB = Colorado River Basin
SSJ = Sacramento-San Joaquin (California)
PNW = Pacific Northwest



Example: How much of change in snow water equivalent is explained by changes in precipitation vs. temperature?



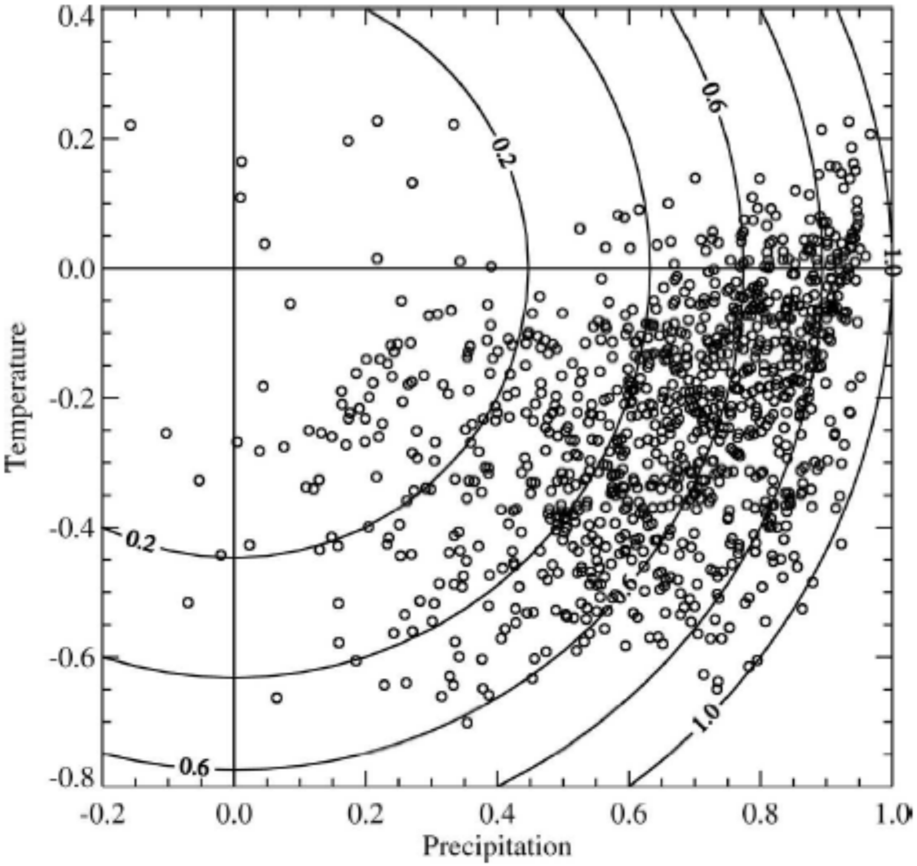


FIG. 3. Each small circle marks the correlations between 1 Apr SWE at one of the 995 snow course locations and the reference time series of Nov–Mar precipitation (x axis) and temperature (y axis). Contours indicate the quantity $(r_T^2 + r_P^2)$, an approximation of the variance explained.

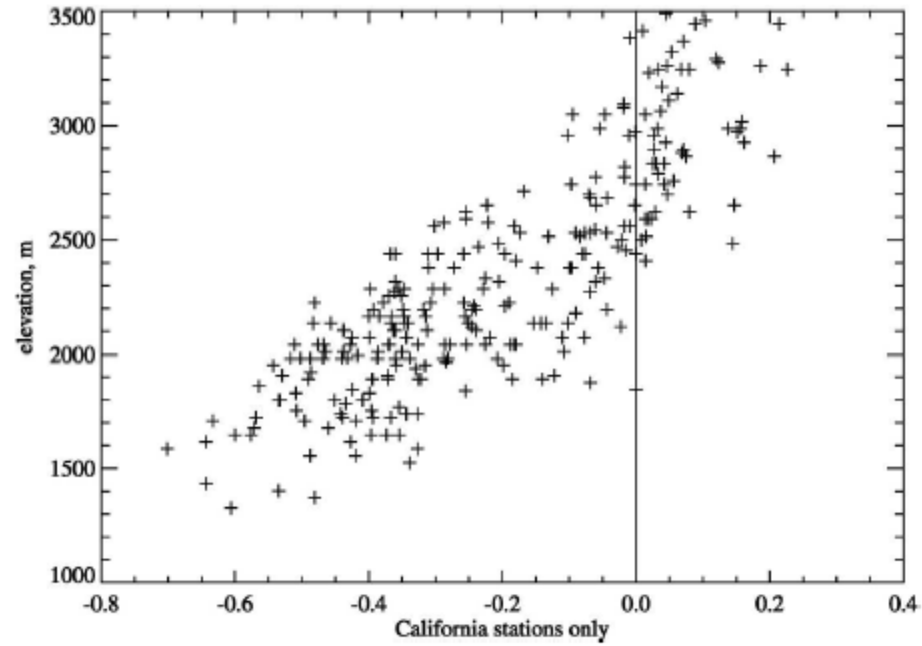
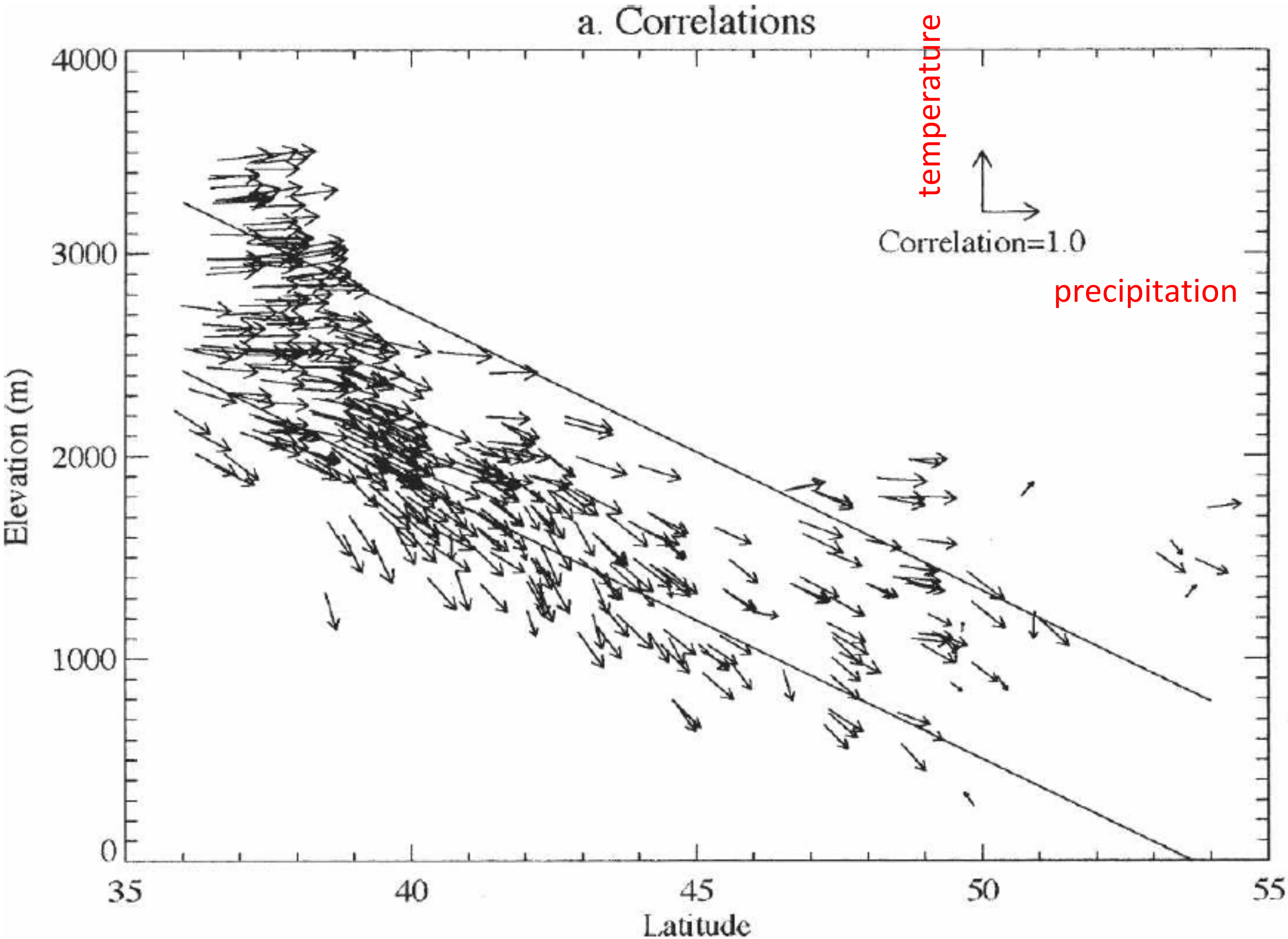


FIG. 4. Correlation between 1 Apr SWE and Nov–Mar temperature at each snow course in California, plotted as a function of snow course elevation.

$\langle S \rangle = a_p \langle P \rangle + a_T \langle T \rangle$ are shown in Fig. 5 for the period 1960–2002. As noted previously for the periods of record 1950–97 (Mote et al. 2005) and 1916–2003 (Hamlet et al. 2005), observed trends in 1 April SWE over the period of record 1960–2002 are also predominantly negative: in fact, the fraction of sites having

Source Mote et al. 2006 (posted under week 3)



Source Mote et al. 2006

Hypothesis Testing:

$$r = \frac{s_{xy}}{\left(\sqrt{\sum(x_i - \bar{x})^2}\right)\left(\sqrt{\sum(y_i - \bar{y})^2}\right)}$$

Null Hypothesis:

$$\rho = 0$$

R and r are used interchangeably here. In Devore, this tests the certainty that your calculated correlation is the true correlation

Test Statistic:

$$t = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

n = Number of samples

Alternate Hypotheses:

$$\rho > 0$$

$$\rho < 0$$

$$\rho \neq 0$$

Rejection Region:

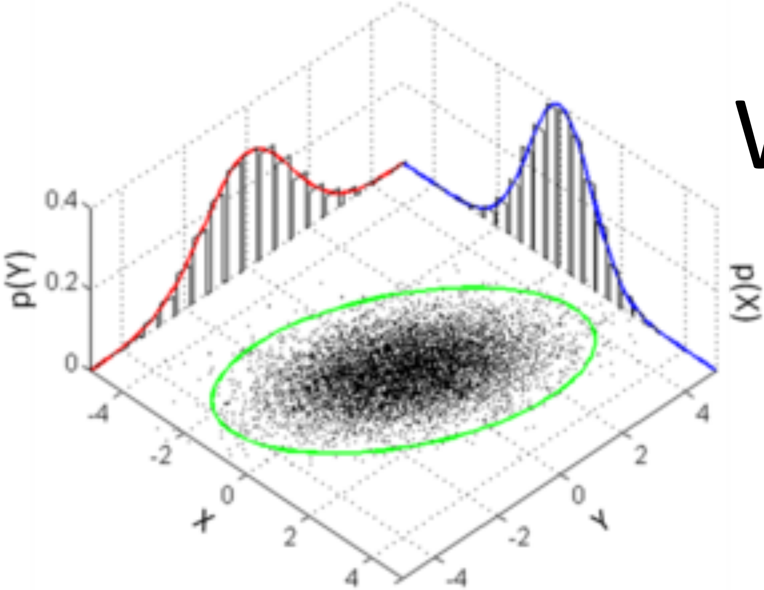
$$t \geq t_{\alpha,n-2}$$

$$t \leq -t_{\alpha,n-2}$$

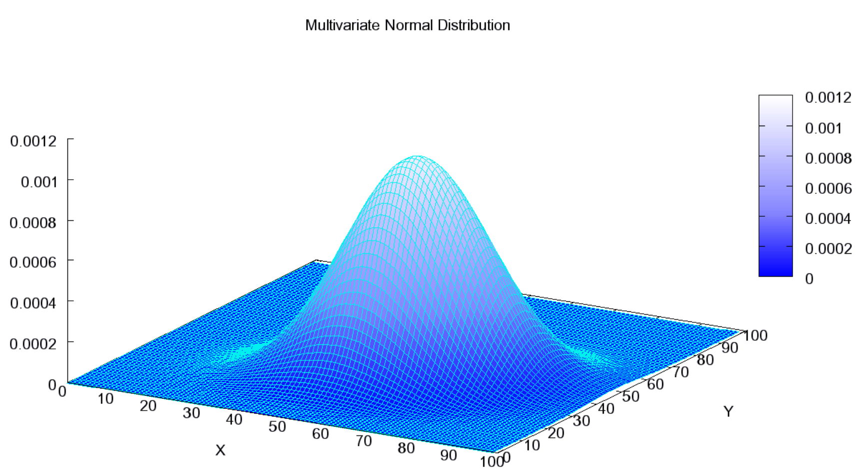
$$t \leq -t_{\frac{\alpha}{2},n-2} \quad OR \quad t \geq t_{\frac{\alpha}{2},n-2}$$

And these are read from the same tables as the t-test, student t distribution.

Why this test statistic?



Examples of bivariate normal (from https://en.wikipedia.org/wiki/Multivariate_normal_distribution)



We're testing for the absence of correlation, but X and Y may be normally distributed and uncorrelated but _not_ independent (recall happy face graphs – anything symmetric).

We assume that both X and Y are random, with a bivariate normal probability distribution (see Section 5.2 in Devore).

After you assume this, you can do a lot of math about drawing numbers from this distribution and chances of getting various correlations among those numbers.

https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient#Testing_using_Student.27s_t-distribution

Hypothesis Testing Part II:

See page 493-494 in Devore
On canvas as
Devore_correlation.pdf

Null Hypothesis:

$$\rho = \rho_0$$

If we say that correlation is a specific number, say you want to be sure you have better than 0.5 correlation

Test Statistic:

$$Let\ V = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right)$$

$$ztest = \frac{V - 1/2 \cdot \ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{1/\sqrt{n-3}}$$

Alternate Hypotheses:

$$\rho > \rho_0$$
$$\rho < \rho_0$$
$$\rho \neq \rho_0$$

Rejection Region:

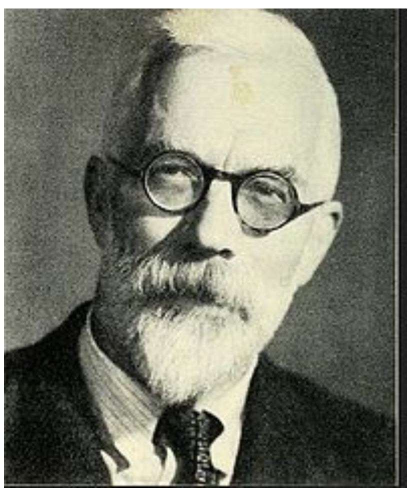
$$ztest > z_\alpha$$
$$ztest < -z_\alpha$$
$$ztest < -z_{\alpha/2}\ OR\ ztest > z_{\alpha/2}$$

Why this formula?

Fisher figured this out in 1915.
https://en.wikipedia.org/wiki/Fisher_transformation

Basically the key idea is to transform the sample correlation coefficient in such a way that the transformed variable would be normally distributed.

You can read these articles if you want to see the math.



$$Let\ V = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right)$$

$$ztest = \frac{V - 1/2 \cdot \ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{1/\sqrt{n-3}}$$

https://en.wikipedia.org/wiki/Fisher_transformation

Regression Models

Homework 3 asks you to develop regression models related to streamflow timeseries.

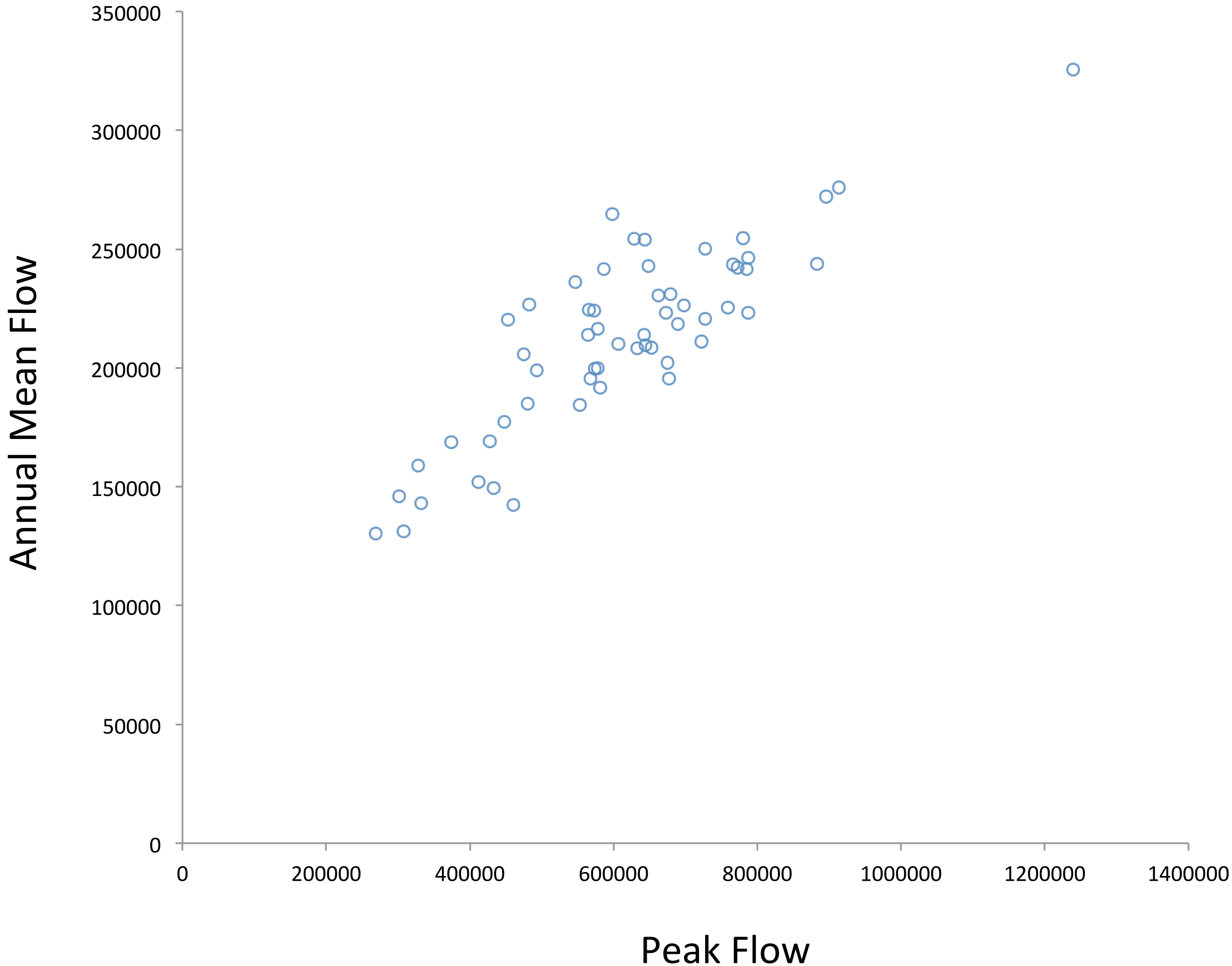
Least Squares Linear Regression:

In this approach we posit a linear relationship between an "independent" or "explanatory" variable x and some "dependent" variable y :

$$y = B_0 + B_1 x$$

The first step in this process is to check whether a linear model approximation is reasonable. A good way to do this is to make a scatter plot of the available data:

Example from Homework 3, Problem 2: Columbia River flow



Fitting of Parameters:

The parameters: $\mathrm{B}_0 \; and \; \mathrm{B}_1$

Are selected so that the sum of the squared errors of the model are minimized for the available data. I.e. minimize:

$$\sum_{i=1}^{n} (y_i - (\mathrm{B}_0 + \mathrm{B}_1 x_i))^2$$

Taking partial derivatives with respect to $\mathrm{B}_0 \; and \; \mathrm{B}_1$ and setting equal to zero yields :

$$n\mathrm{B}_0 + \left(\sum_{i=1}^{n} x_i\right)\mathrm{B}_1 = \left(\sum_{i=1}^{n} y_i\right)$$

$$\left(\sum_{i=1}^{n} x_i\right)\mathrm{B}_0 + \left(\sum_{i=1}^{n} x_i^2\right)\mathrm{B}_1 = \left(\sum_{i=1}^{n} x_i \, y_i\right)$$

Solving for $\mathrm{B}_0\ and\ \mathrm{B}_1$ yields:

$$\mathrm{B}_1 = \frac{n\left(\sum_{i=1}^{n} x_i\, y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\mathrm{B}_0 = \frac{\left(\sum_{i=1}^{n} y_i\right) - \mathrm{B}_1\left(\sum_{i=1}^{n} x_i\right)}{n} = \bar{y} - \mathrm{B}_1\bar{x}$$

$$Let\ \hat{y}_i = \mathrm{B}_0 + \mathrm{B}_1 x_i$$

Then the quantity $(y_i - \widehat{y_i})$ is called the "*i*th residual" .

Let:

$$SSE = Sum\ of\ Squared\ Errors$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\sigma^2 = s^2 = \frac{SSE}{(n-2)}$$

$$\sigma = \sqrt{\frac{SSE}{(n-2)}}$$

s is also called the “standard error” of the regression model.

$$SST = Total\ Sum\ of\ Squares$$

$$Let\ SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

How much variance is there about the mean.

$$R^2 = 1 - \frac{SSE}{SST}$$

R^2 is often described as the fraction of the variance explained by the model. If the model is no better than predicting the mean, then the variance explained would be zero. A perfect model (i.e. $SSE = 0$) is said to explain 100% of the variance.

Note similarity here to the ANOVA formulation. You can often see people using ANOVA analysis to discuss the variance explained by a specific grouping or classification. ANOVA is sometimes considered a special case of linear regression.

Confidence Bounds on Regression Parameters:

The variance of the regression parameter $\hat{B}_1$ is a function of the standard error *and* the "spread" of the x values.

$$s_{B_1}{}^2 = \frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

And $\dfrac{(\hat{B}_1 - \mathrm{B}_1)}{s_{B_1}}$ is T distributed with n-2 degrees of freedom.

So a confidence interval for B_1 is: $\hat{B}_1 \pm t_{\frac{\alpha}{2},n-2} \cdot s_{B_1}$

Hypothesis test for the estimator $\widehat{B}_1$

Asking, "Is there really a slope to my regression line?"

Null Hypothesis: $\widehat{B}_1 = \mathrm{B}_1$

α = Probability of a Type I error, number of degrees of freedom = (n-2)

Test statistic: $$t = \frac{(\widehat{B}_1 - \mathrm{B}_1)}{s_{B_1}}$$

Alternate Hypothesis:	Rejection Region:
$\widehat{B}_1 > \mathrm{B}_1$	$t \geq t_{\alpha,n-2}$
$\widehat{B}_1 < \mathrm{B}_1$	$t \leq -t_{\alpha,n-2}$
$\widehat{B}_1 \neq \mathrm{B}_1$	$t \leq -t_{\frac{\alpha}{2},n-2} \; OR \; t \geq t_{\frac{\alpha}{2},n-2}$

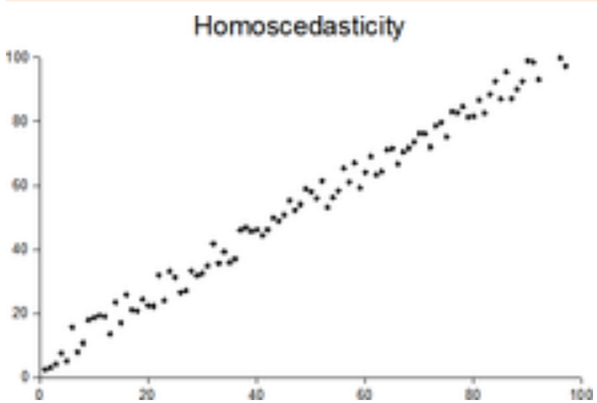
Estimating the Trend Using a Least Squares Linear Model:

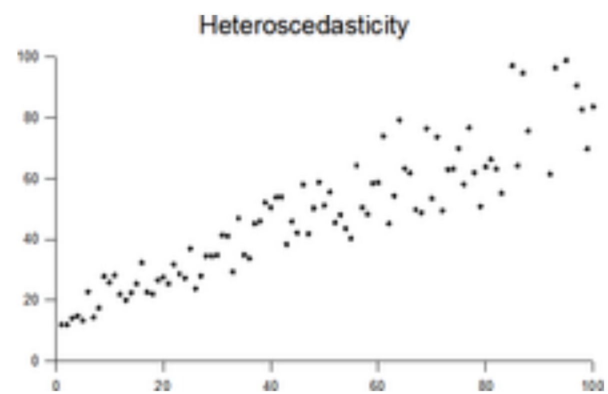
Some conditions that should be met for good results (see Helsel and Hirsch for more)

- Data should not be strongly auto correlated (more on this next week),
- There shouldn't be any dramatic expansion in the variance over time.
- A linear model should fit reasonably well (use a scatter plot to confirm)
- The residuals for the linear model should be approximately normally distributed and shouldn't have large trends in them (plot these to get a sense of whether there are problems).

Homoscedasticity: random variables in a sequence have the same finite variance.

Heteroscedasticity: subpopulations have different variance from others.





Graphs from wikipedia: https://en.wikipedia.org/wiki/Heteroscedasticity

Estimating the Trend Using a Least Squares Linear Model:

Some conditions that should be met for good results (see Helsel and Hirsch for more)

•Data should not be strongly auto correlated,
•There shouldn't be any dramatic expansion in the variance over time.
•A linear model should fit reasonably well (use a scatter plot to confirm)
•The residuals for the linear model should be approximately normally distributed and shouldn't have large trends in them (plot these to get a sense of whether there are problems).

Procedures:

•Calculate B_1 (the trend) in the normal manner. (What are the units?)

•Use hypothesis tests on B_1 to see whether the trend is significantly different from 0 (i.e. no trend).

•Use the confidence interval around the estimate of B_1 to express the uncertainty in the trend.

Using the LINEST Function in Excel:

Excel Help

Search

- **stats** Optional. A logical value specifying whether to return additional regression statistics.
  - If ***stats*** is TRUE, **LINEST** returns the additional regression statistics; as a result, the returned array is **{mn,mn-1,...,m1,b;sen,sen-1,...,se1,seb;r2,sey;F,df;ssreg,ssresid}**.
  - If ***stats*** is FALSE or omitted, **LINEST** returns only the m-coefficients and the constant b.

The additional regression statistics are as follows.

STATISTIC	DESCRIPTION
se1,se2,...,sen	The standard error values for the coefficients m1,m2,...,mn.
seb	The standard error value for the constant b (seb = #N/A when ***const*** is FALSE).
r2	The coefficient of determination. Compares estimated and actual y-values, and ranges in value from 0 to 1. If it is 1, there is a perfect correlation in the sample — there is no difference between the estimated y-value and the actual y-value. At the other extreme, if the coefficient of determination is 0, the regression equation is not helpful in predicting a y-value. For information about how r2 is calculated, see "Remarks," later in this topic.
sey	The standard error for the y estimate.
F	The F statistic, or the F-observed value. Use the F statistic to determine whether the observed relationship between the dependent and independent variables occurs by chance.
df	The degrees of freedom. Use the degrees of freedom to help you find F-critical values in a statistical table. Compare the values you find in the table to the F statistic returned by **LINEST** to determine a confidence level for the model. For information about how df is calculated, see "Remarks," later in this topic. Example 4 shows use of F and df.
ssreg	The regression sum of squares.
ssresid	The residual sum of squares. For information about how ssreg and ssresid are calculated, see "Remarks," later in this topic.

The following illustration shows the order in which the additional regression statistics are returned.

	A	B	C	D	E	F
1	m_n	m_{n-1}	...	m_2	m_1	b
2	se_n	se_{n-1}	...	se_2	se_1	se_b
3	r_2	se_y				
4	F	d_f				
5	ss_{reg}	ss_{resid}				

$$seb1 = s_{B_1}$$
$$sey = s$$
$$ssresid = SSE$$
$$ssreg = SST - SSE$$

Note that LINEST function produces a table of output, so select a group of cells before typing in the formula and complete your formula entry as an array (ctrl-shift-enter).

Using the LINEST Function in Excel:

Coefficients of the regression, starting from highest

Standard error of each of those coefficients

	A	B	C	D	E	F
1	m_n	m_{n-1}	. . .	m_2	m_1	b
2	se_n	se_{n-1}	. . .	se_2	se_1	se_b
3	r_2	se_y				
4	F	d_f				
5	ss_{reg}	ss_{resid}				

$$seb1 = s_{B_1}$$
$$sey = s$$
$$ssresid = SSE$$
$$ssreg = SST - SSE$$

Standard error in y, or standard error of the model

$$sey = s$$

R^2 – coefficient of determination

Sum of squared errors

Note that LINEST function produces a table of output, so select a group of cells before typing in the formula and complete your formula entry as an array (ctrl-shift-enter).

In a Mac, type ⌘+RETURN

LINEST Examples:

See example spreadsheet LINEST_examples.xlsx in tools folder on website

Constructing a Confidence Interval for the Predicted Values of Y:
(see Regression_conf_intervals_Devore_p483.pdf or
https://www.ma.utexas.edu/users/mks/statmistakes/CIvsPI.html)

$$Let\ x^*\ be\ a\ particular\ value\ of\ x$$

$$Let\ y^* = \mathrm{B}_0 + \mathrm{B}_1 x^*$$

$$Var[Y - (\mathrm{B}_0 + \mathrm{B}_1 x^*)] = Var(Y) + Var(\mathrm{B}_0 + \mathrm{B}_1 x^*)$$

$$Var(Y)\ is\ the\ variance\ of\ Y\ relative\ to\ the\ model$$

How well did your model actually fit the data

(That is, the standard error of the model σ squared.)

$$and\ Var(\mathrm{B}_0 + \mathrm{B}_1 x^*)$$
$$is\ the\ variance\ of\ the\ model\ prediction.$$

How far away from your data mean are you

Let:

How well did your model actually fit the data

$$SSE = Sum\ of\ Squared\ Errors$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\sigma^2 = s^2 = \frac{SSE}{(n-2)}$$

$$\sigma = \sqrt{\frac{SSE}{(n-2)}}$$

s is also called the “standard error” of the regression model.

The combined variance of the error of prediction at x^* can be shown to be:

$$\sigma^2_{E_p}(x^*) = var\,(y - y^*) =$$

$$= s^2\left[1 + \frac{1}{n} + \frac{n(x^* - \bar{x})^2}{n\sum_{i=1}^{n} x_i^2 + \left(\sum_{i=1}^{n} x_i\right)^2}\right]$$

Note: Xbar and Xi refer to the ORIGINAL data used to make the model. S is the original standard error.
And the statistic:

$$T = \frac{(y - y^*)}{\sigma_{E_p}(x^*)}$$

has a t distribution with n-2 degrees of freedom. Note: TINV in excel looks up a t-table. Also, check out tinv.m in matlab (built in function that looks up the t-table).

Thus a $(1-\alpha)$ confidence interval for

y *at an arbitrary value of* x^* *is*:

$$y^* \pm t_{\frac{\alpha}{2},n-2} \cdot \sigma_{E_p}(x^*)$$

Note that the uncertainty is a function of x^* *and the farther away from* $\bar{x}$ *we find ourselves the larger the uncertainty in the prediction of y!*

(Key thing to remember, these are not constant and vary with the location you want to predict.)

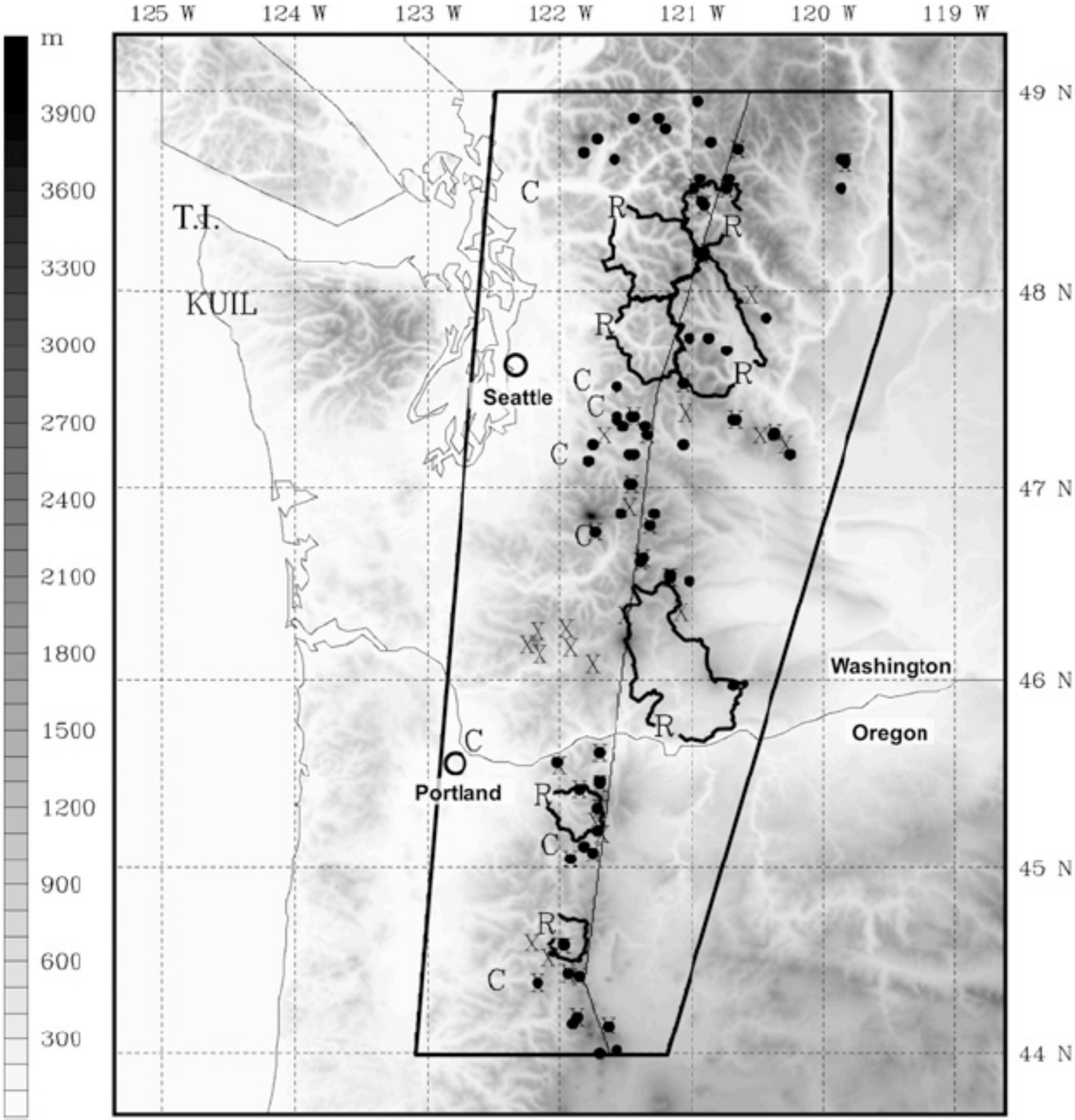
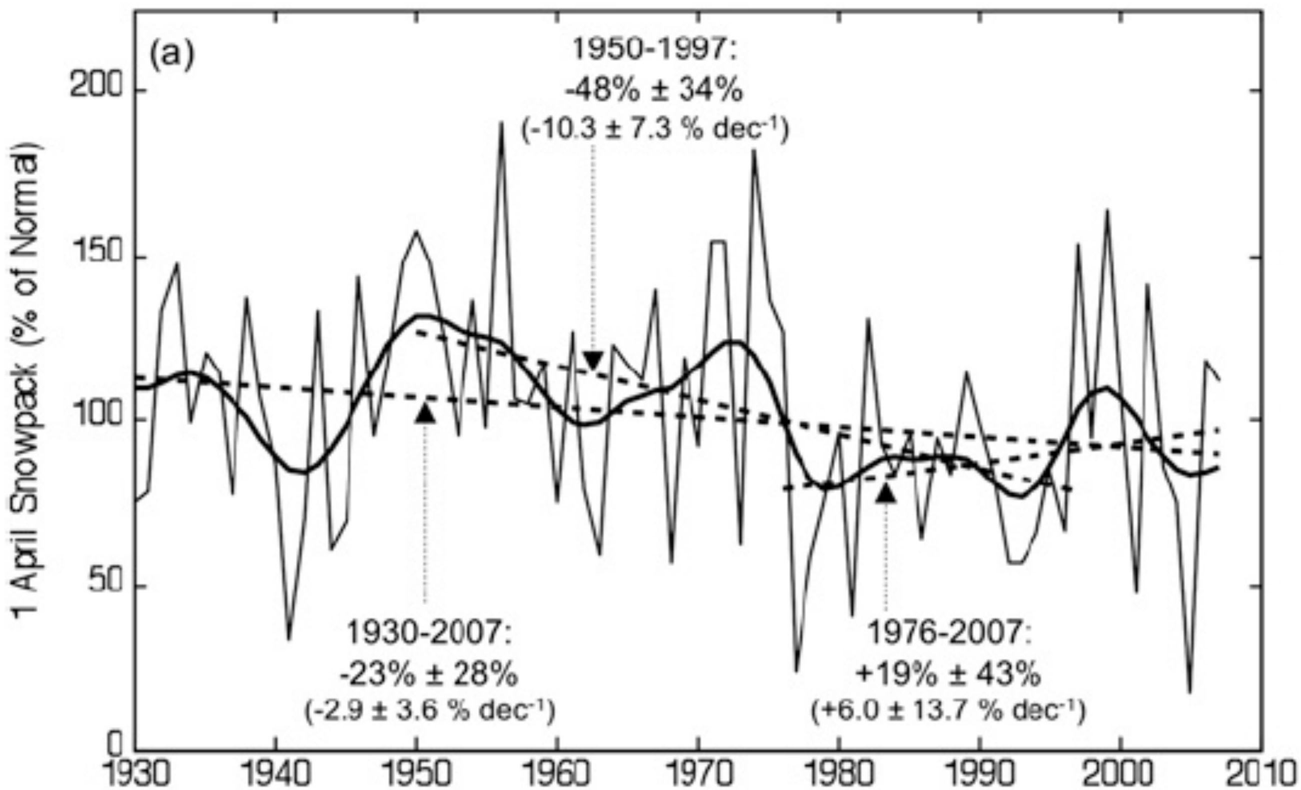


FIG. 1. Map of study area. Heavy solid polygon defines "Cascade Mountains" for the purposes of this study. The thin solid line di-

Stoelinga et al. 2010: Cascades Snowpack trends

What is shown here: Confidence values in predicted snow values based on a linear fit? Or uncertainties in the fitted slope of the line?



Snowpack trends for different periods: Note importance of the confidence intervals in those trends (from Stoelinga et al. 2010)

Non-Parametric Quantile Regression

Non-Parametric Regression:

Non-parametric regression approaches have many advantages:

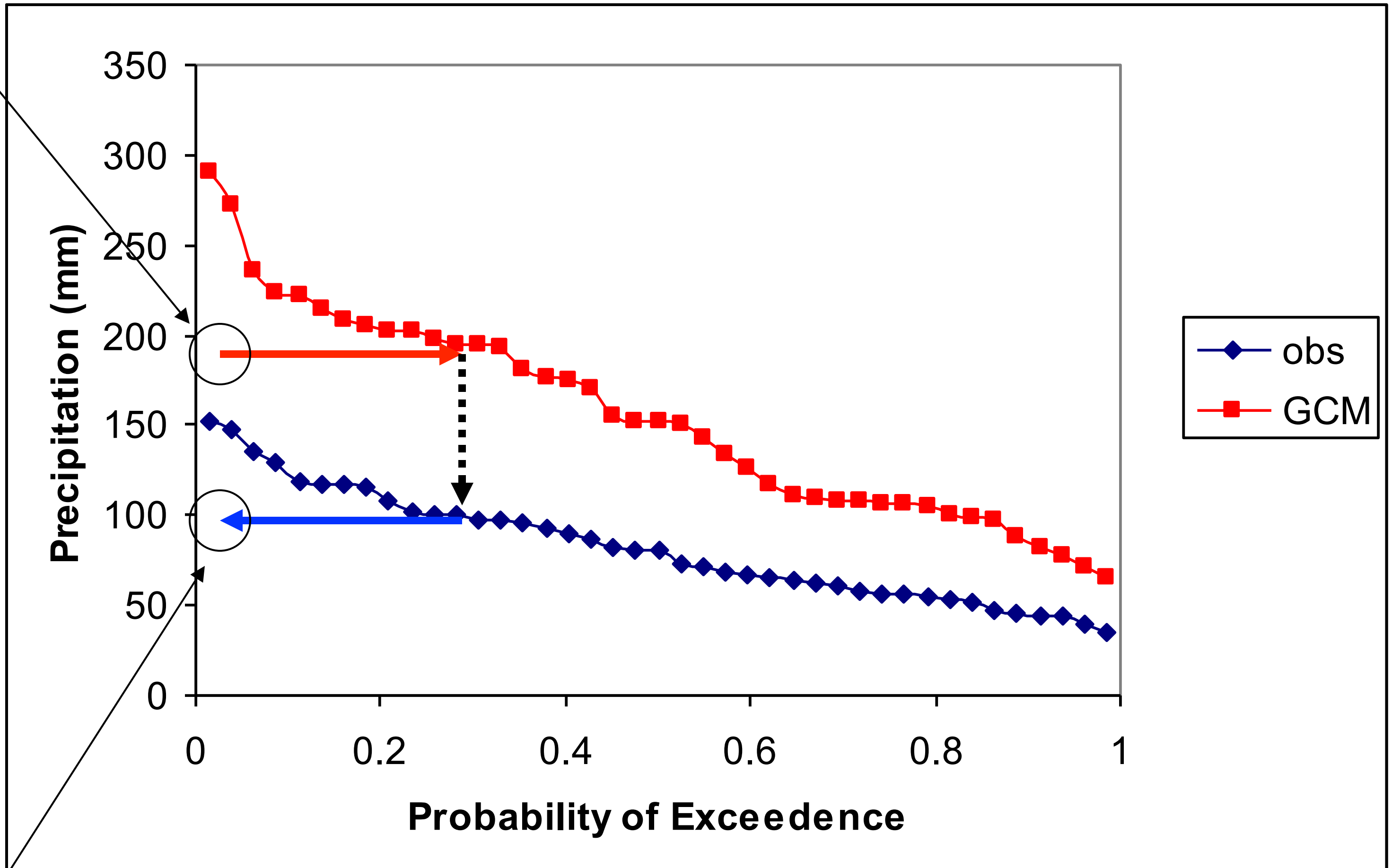
- Do not require that the underlying probability distributions are known or have any particular form.

- A linear relationship between the two variables is not required.

- The time series of the data need not be the same (or even from the same time period) in the explanatory and dependent variables . That is, paired data is not required (although in many cases it is desirable).

We presume that relative ranking and frequencies of events are correct, even if actual values don't match up in a linear way.

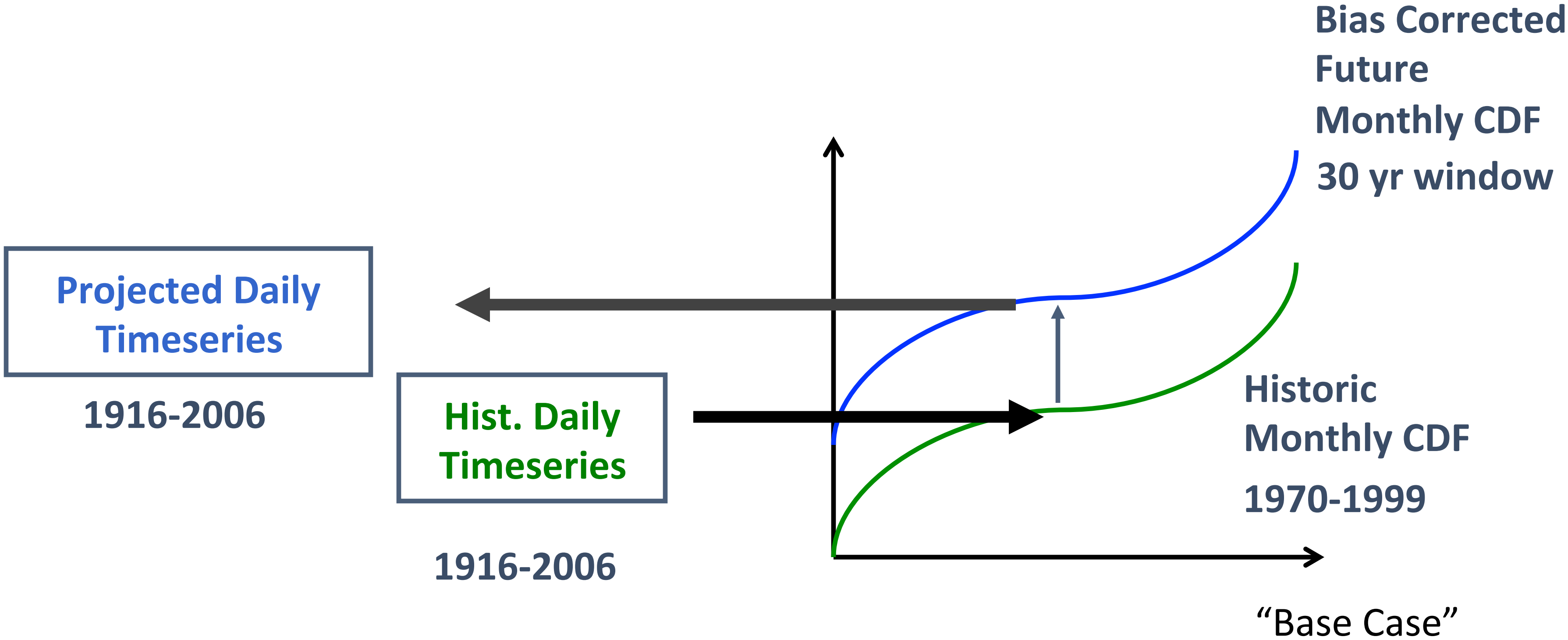
GCM Input = 190



Bias Corrected Output = 100

Hybrid Downscaling Method

- Performed for each VIC grid cell:



- Used to correct hydro models in places where groundwater important
- Can be biased in absolute values but still have good change signals

Excel Example of Quantile Mapping Process:

See example spreadsheet quantile_regression. Xlsx

In tools folder

We will also work with this in today's lab.