

# Math is Music – Stats is Literature

*Or why are there no six year old novelists?*

**Dick De Veaux - Williams College**

Thanks also to Paul Velleman,

**Cornell University**



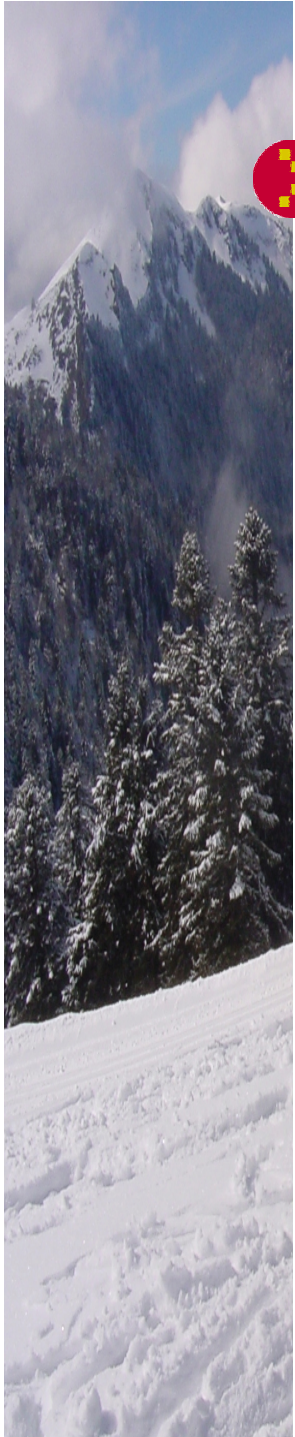
# Prodigies

- **Math, music, chess**
  - Gauss, Pascal
  - Mozart, Schubert, Mendelssohn
  - Bobby Fischer
- **Why these three areas?**
- **Each creates its own world with its own set of rules**
  - There is no “experience” required
  - Once you know the rules, you are free to create anything



# Prodigies in Literature

- **Mary Wollstonecraft Shelley**
  - Age 19
  - Created Frankenstein, an imaginary creature
- **Others?**
- **Why?**
  - Literature is about the world, not about rules. It deals with life's experience and the wisdom we develop over time.



# Statistics – What do students find hard?

- “Understood the material in class, but found it hard to do the homework”
- “Should be more like a math course, with everything laid out beforehand”
- “More problems in class should be like the HW and tests”



# What is “easy”?

- **The math part**
  - Give them the formula, they can get the answer with some training
- **The hard part**
  - Putting it all together
    - Real world
    - Experience
    - Methods

# What's "Hard"? -- Example

JMP - cup98lrn.4.28

File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help

cup98lrn.4.28

cup98lrn.4.28

	ODATEDW	OSOURCE	TCODE	STATE	ZIP	MAILCODE	PVASTATE	DOB	NOEXCH	RECINHSE	RECP3	RECPGVG	RECSWEEP	MC
1404	9201	MCC	0	FL	32771	N	N	5010	0	N	N	N	N	XX
1405	9201	HOS	0	CA	94112	N	N	4105	0	X	N	N	N	XX
1406	9201	SYN	0	TX	76401	N	N	3407	0	N	N	N	N	XX
1407	9501	ARG	1	CA	95726	N	N	4501	0	N	N	N	N	XX
1408	9001	MON	1	TX	76570	N	N	1001	0	X	N	N	N	XX
1409	8701	LIS	1	CA	91770	N	N	5901	0	N	N	N	N	XX
1410	9101	L01	1	VM	53142	N	N	5601	0	N	N	N	N	XX
1411	9601	MCO	2	FL	32408	N	N	2301	0	N	N	N	N	XX
1412	9601	MCO	0	CA	93309	B	N	0	0	N	N	N	N	XX
1413	8601	VKB	0	UT	84103	B	N	4601	0	N	N	N	N	XX
1414	9601	DCD	1	TN	38125	N	N	5001	0	N	N	N	N	XX
1415	9101	LHJ	0	GA	31544	N	N	4507	0	X	N	N	N	XX
1416	8901	SES	1	CO	80816	N	N	0	0	N	N	N	N	XX
1417	8601	MBC	2	VM	53006	N	N	907	0	N	N	N	N	XX
1418	9401	MSD	1	MS	38663	N	N	1901	0	N	N	N	N	XX
1419	8601	GUR	0	ID	83642	N	N	3401	0	X	N	N	N	XX
1420	9401	HOS	0	GA	31328	N	N	5401	0	N	N	N	N	XX
1421	9001	AML	0	CA	91770	N	N	4510	0	N	N	N	N	XX
1422	8701	LIS	1	OR	97523	N	N	0	0	N	N	N	N	XX
1423	9501	KIP	0	IL	61821	N	N	0	0	N	N	N	N	XX
1424	9501	USB	0	NV	89104	N	N	3001	0	N	N	N	N	XX
1425	9101	DUR	0	AL	36525	N	N	0	0	N	N	N	N	XX
1426	9301	ALZ	0	IL	60645	N	P	908	0	N	N	N	X	XX
1427	8801	LIS	28	TX	77090	N	N	3901	0	N	N	N	N	XX
1428	9301	AGR	0	VM	53402	N	N	3401	0	N	N	N	N	XX
1429	8601	ENQ	2	IN	46561	N	N	1306	0	N	N	N	N	XX
1430	9301	SYN	0	MO	64779	N	N	4301	0	N	N	N	N	XX

Columns (379/1)

- ODATEDW
- OSOURCE
- TCODE
- STATE
- ZIP
- MAILCODE
- PVASTATE
- DOB
- NOEXCH
- RECINHSE
- RECP3
- RECPGVG
- RECSWEEP
- MDMAUD
- DOMAIN
- CLUSTER
- AGEFLAG
- HOMEOWNR
- NUMCHLD
- INCOME
- GENDER

Rows

All Rows 94649  
Selected 1  
Excluded 0

Method

Oneway Anova

Summary of Fit

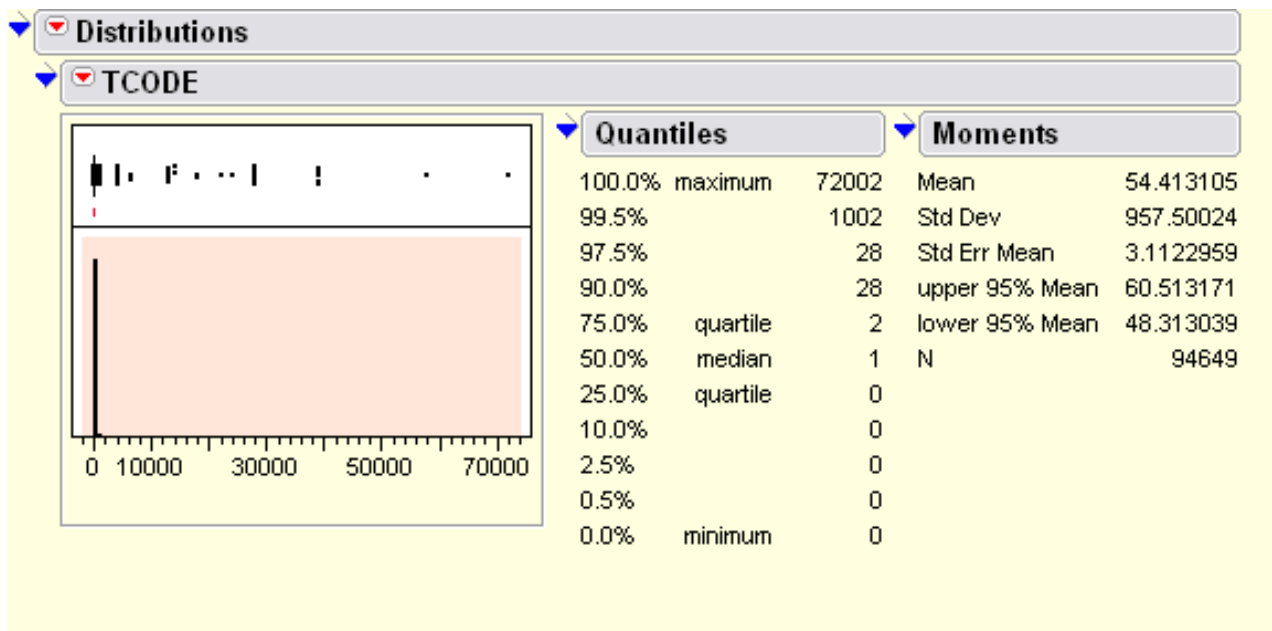
Std Err Mean 3.1122959  
upper 95% Mean 60.513171  
lower 95% Mean 48.313039  
N 94649

September 18, 2004

AWL Workshop -- HACC

6

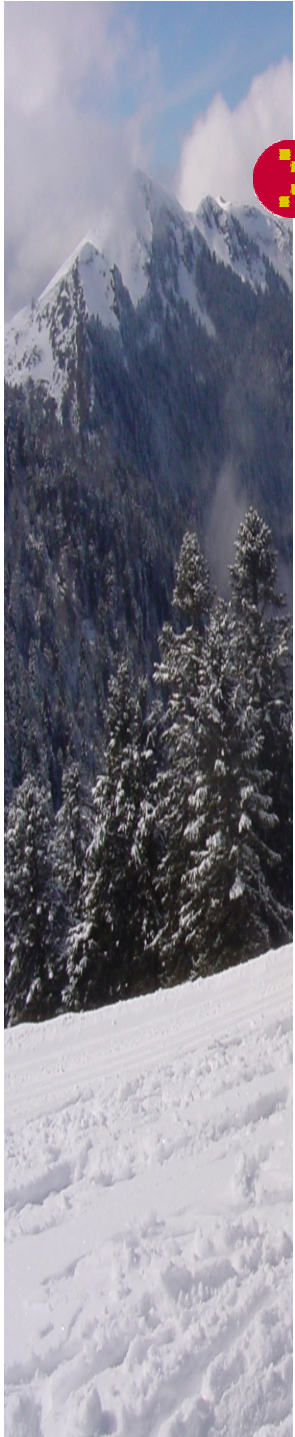
# T-Code





# What does it mean?

T-Code	Title						
0	_	16	DEAN	48	CORPORAL	109	LIC.
1	MR.	17	JUDGE	50	ELDER	111	SA.
1001	MESSRS.	17002	JUDGE & MRS.	56	MAYOR	114	DA.
1002	MR. & MRS.	18	MAJOR	59002	LIEUTENANT & MRS.	116	SR.
2	MRS.	18002	MAJOR & MRS.	62	LORD	117	SRTA.
2002	MESDAMES	19	SENATOR	63	CARDINAL	118	SRTA.
3	MISS	20	GOVERNOR	64	FRIEND	120	YOUR MAJESTY
3003	MISSSES	21002	SERGEANT & MRS.	65	FRIENDS	122	HIS HIGHNESS
4	DR.	22002	COLNEL & MRS.	68	ARCHDEACON	123	HER HIGHNESS
4002	DR. & MRS.	24	LIEUTENANT	69	CANON	124	COUNT
4004	DOCTORS	26	MONSIGNOR	70	BISHOP	125	LADY
5	MADAME	27	REVEREND	72002	REVEREND & MRS.	126	PRINCE
6	SERGEANT	28	MS.	73	PASTOR	127	PRINCESS
9	RABBI	28028	MSS.	75	ARCHBISHOP	128	CHIEF
10	PROFESSOR	29	BISHOP	85	SPECIALIST	129	BARON
10002	PROFESSOR & MRS.	31	AMBASSADOR	87	PRIVATE	130	SHEIK
10010	PROFESSORS	31002	AMBASSADOR & MRS	89	SEAMAN	131	PRINCE AND PRINCESS
11	ADMIRAL	33	CANTOR	90	AIRMAN	132	YOUR IMPERIAL MAJEST
11002	ADMIRAL & MRS.	36	BROTHER	91	JUSTICE	135	M. ET MME.
12	GENERAL	37	SIR	92	MR. JUSTICE	210	PROF.
12002	GENERAL & MRS.	38	COMMODORE	100	M.		
13	COLONEL	40	FATHER	103	MLLE.		
13002	COLONEL & MRS.	42	SISTER	104	CHANCELLOR		
14	CAPTAIN	43	PRESIDENT	106	REPRESENTATIVE		
14002	CAPTAIN & MRS.	44	MASTER	107	SECRETARY		
15	COMMANDER	46	MOTHER	108	LT. GOVERNOR		
15002	COMMANDER & MRS.	47	CHAPLAIN				



# What's Hard?

## Five Unnatural Acts

- Think *Critically*
- Be *Skeptical*
- Focus not on what we know, but on what we *don't know*
- Think first about *Variation*
- Think clearly about *Conditioning* and *Rare* events



# Statistics is Unnatural and Subversive

- **We ask students to**
  - Question the data
  - Examine the assumptions
  - Reject the null hypothesis
- **Have they done this in “math” class?**
- **Convincing them to be subversive may be easier than you think**



# Think Critically

- Challenge the data's credentials.
- Look for bias.
- Know what we want to know.
  - What's the QUESTION?
- Look for Lurking variables.
- Check Assumptions and Conditions.

**Critical thinking requires creativity. You must think about things that are not in front of you and imagine ways in which things *might* have gone wrong.**



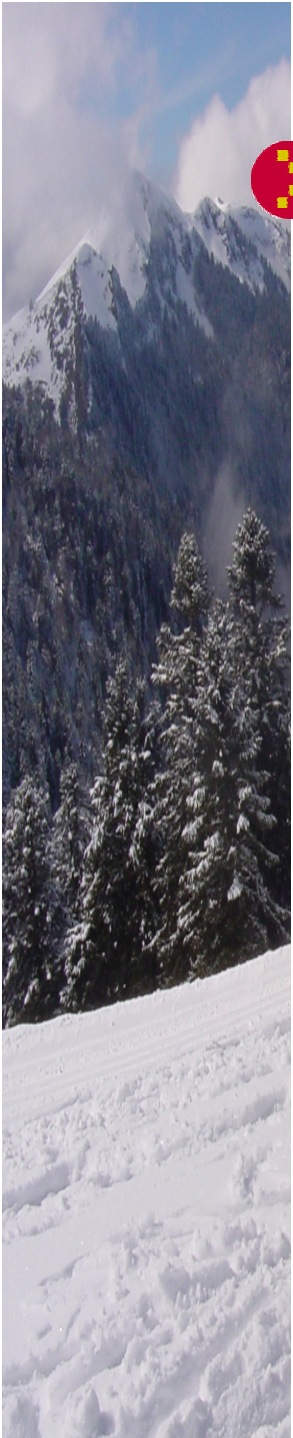
# Be Skeptical

- Be cautious about making claims based on data.
- “*Trust every analysis, but plot the residuals.*”
  - Skeptical statisticians expect the unexpected, so we go looking for it.
- **SHOW** that the analysis is appropriate

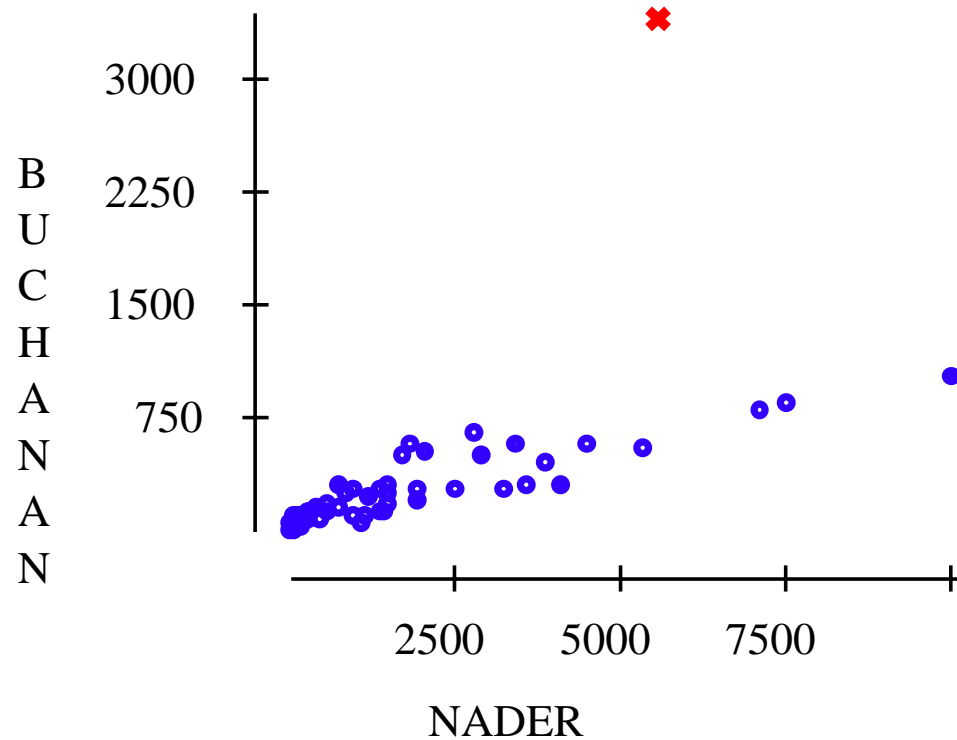


# Ancient History

- The vote in the 2000 Presidential election for Buchanan and the vote for Nader, (the two principal alternatives to Bush and Gore), has a correlation of 0.65 over the counties of Florida.
- Ask:
  - Is the relationship *linear*?
  - Is the data set homogeneous or are there *subgroups*?
  - Are there any *outliers*?



# Plot the Data



Without Palm Beach county and its “butterfly ballot”, the correlation is 0.91.



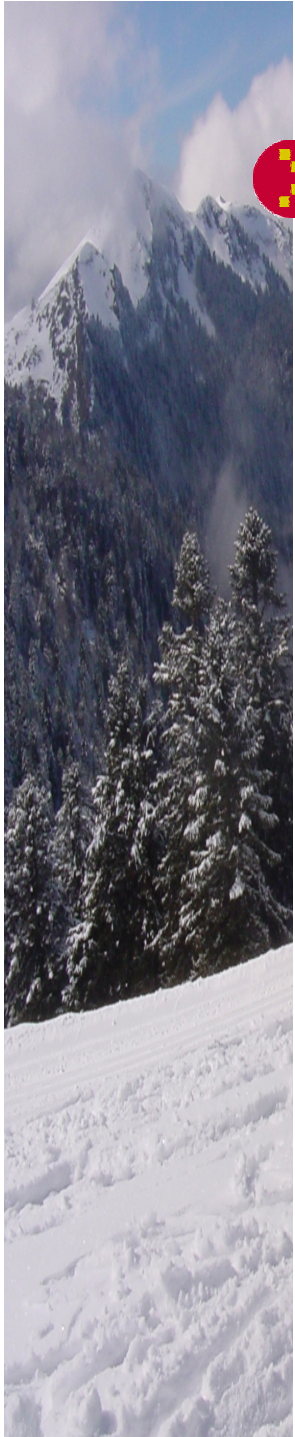
# Hypothesis Testing

- **Skepticism formalized**
- **The null hypothesis is a skeptical claim about the data**
- **It's unnatural to show the opposite**



# Critical Thinking and Skepticism

- **Critical thinking** is open-ended questioning of the data's credentials.
  - We wonder whether the data are competent to tell us what we want to know.
- **Skepticism** questions whether what the data appear to be telling us is the whole truth.



# Focus on What We Don't Know

- In most science and math courses, we focus on what we know
- Statisticians are a bit perverse



# Confidence Intervals

- We don't say "*The mean is 31.2*".
- We don't say "*The mean is probably 31.2*".
- We don't say "*The mean is close to 31.2*".
- All we can manage is
  - "***The mean is close to 31.2.... Probably***
  - ***(and, in fact, I'm willing to admit I may be wrong and to spend the effort to give you a whole interval of plausible values and then to spend extra effort to estimate how likely it is that even that interval is wrong.)***"



# All Models are Wrong...

**George Box:**

**“All models are wrong... but some are useful”**

**“Statisticians, like artists, have the bad habit of falling in love with their models”**

**But, statisticians love models--*because they are wrong.***

**What do we focus on?**

**residuals!**

**what the model fails to account for**



# Variation

- **Students find it easier to think about values rather than variation, but**

*Statistics is about Variation*



# Example

- **A town has two hospitals**
  - Large hospital about 100 babies a day
  - Smaller hospitals about 15 babies a day
- **Over the course of the year, which hospital (if either) would probably have more days in which more than 60% of the babies born are male?**



# The Standard Deviation is the Statistician's Ruler

- Most of the inference seen in the introductory course compares a statistic to its standard deviation to see whether it is “big”.
- This idea carries into advanced methods as well.



# Thinking about Conditional Events

- **This is just plain hard.**
- **It is easy to show that we don't naturally think clearly about conditional probabilities.**
- **But we must for rational decision making.**



# Linda

(Tversky & Kahneman)

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and she participated in antinuclear demonstrations.



## Order these in order of Likelihood

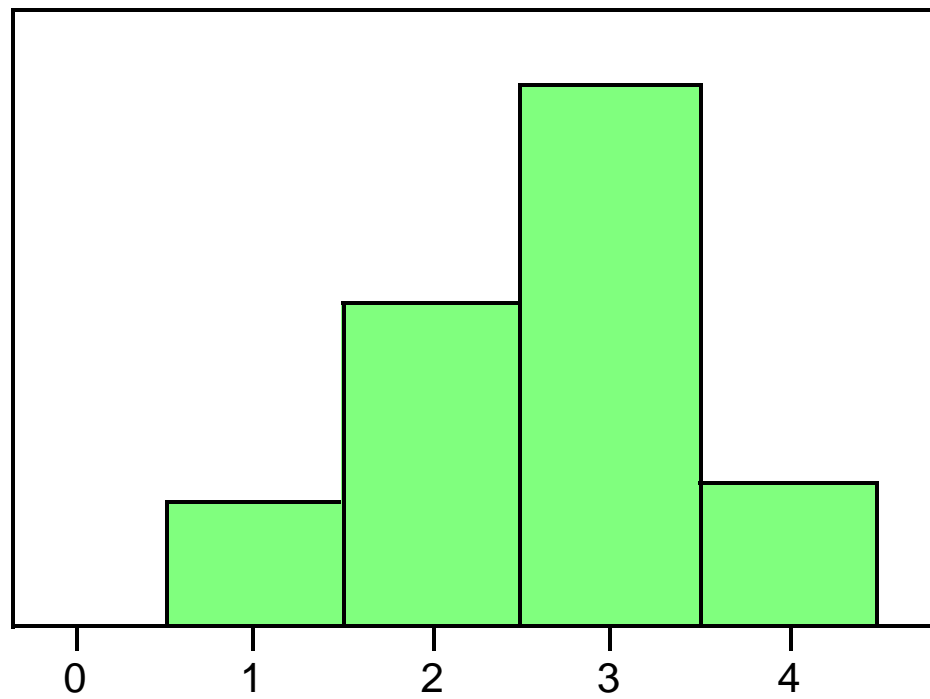
- a) Linda is a teacher in an elementary school
- b) Linda works in a bookstore and takes yoga classes.
- c) Linda is active in the feminist movement.
- d) Linda is a psychiatric social worker
- e) Linda is a member of the League of Women Voters.
- f) Linda is a bank teller.
- g) Linda is an insurance salesperson.
- h) Linda is a bank teller who is active in the feminist movement.



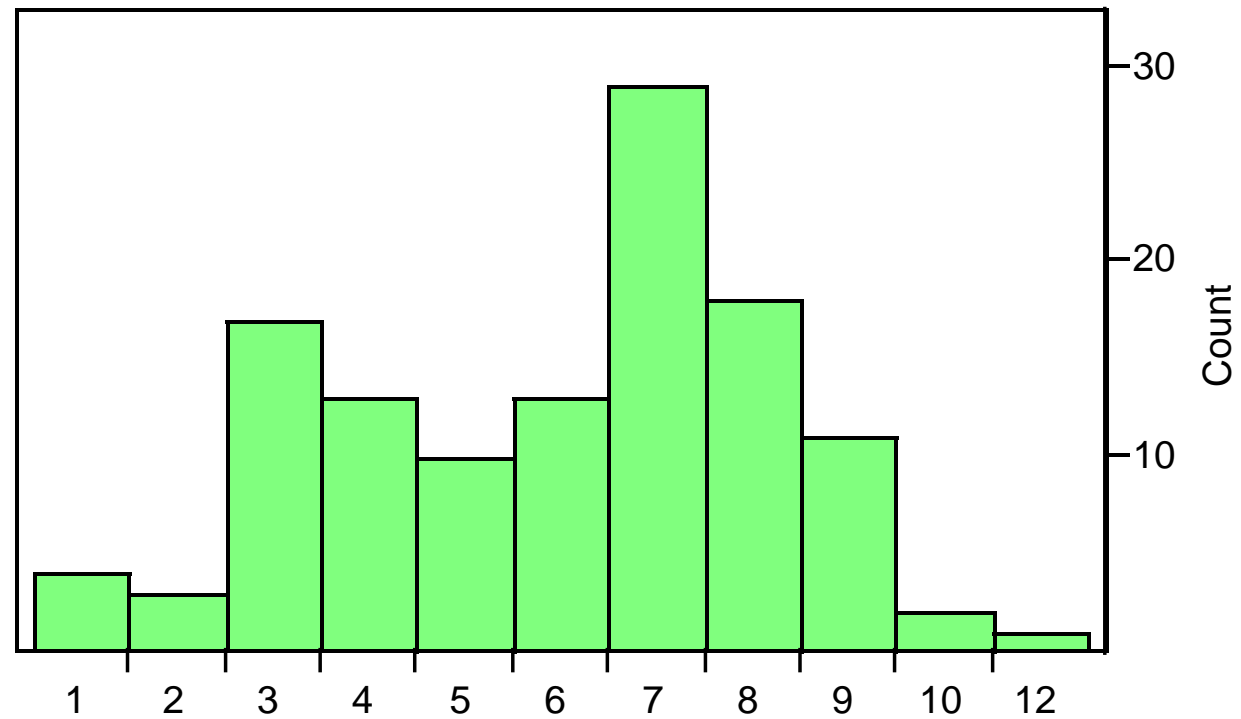
# Pick a number at Random

1 2 3 4

# Random?



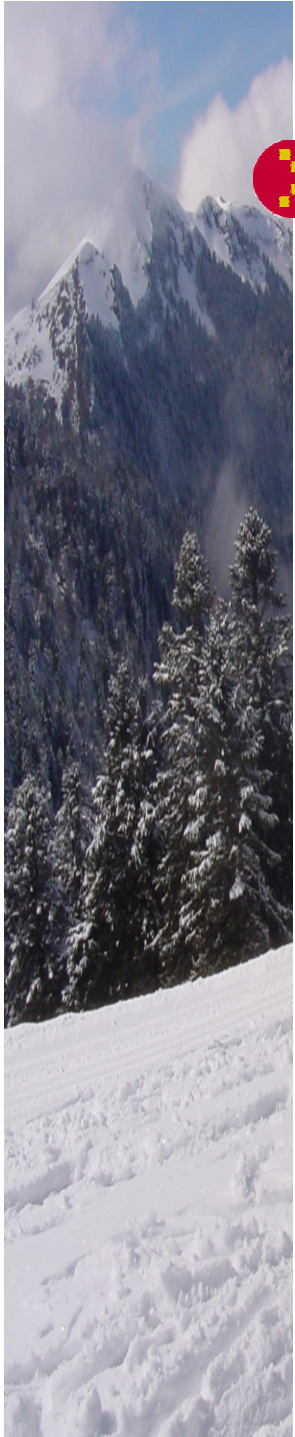
# Random II



September 18, 2004

AWL Workshop -- HACC

28



# Is Statistical Thinking Unnatural?

- We haven't evolved to be Statisticians.
- Our students who think Statistics is an unnatural subject are right. This isn't how humans think naturally.
- But it is how humans think rationally. And it is how scientists think. This is the way we must think if we are to make progress in understanding how the world works and, for that matter, how we ourselves work.



# How can we help?

- **Give them an outline for putting the real world into a framework**
  - **What's the problem?**
    - The W's
    - The model
    - The method
  - **What are the mechanics?**
  - **What have we learned?**



# Think – Show -- Tell

- **THINK:** What techniques apply?
- **SHOW:** Mechanics – how to do it.
- **TELL:** Explain what you learned.

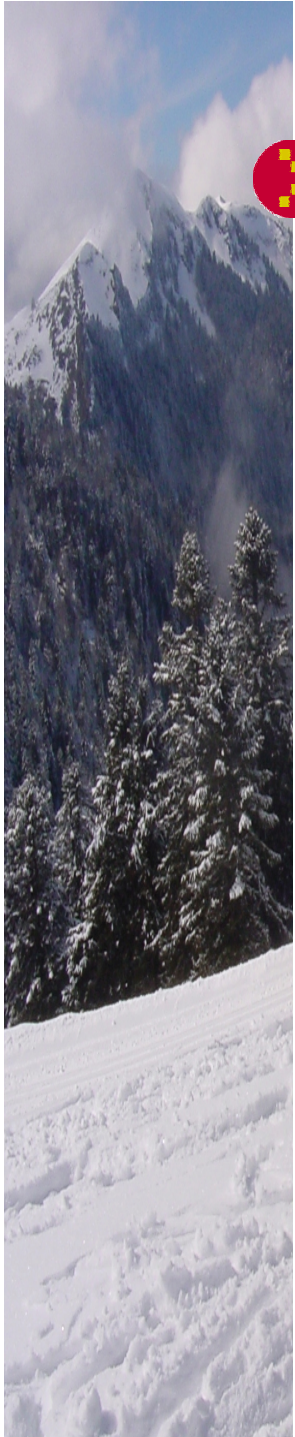


# Think

September 18, 2004

AWL Workshop -- HACC

32



# The Three Rules of Data Analysis

## I. Make a Picture

it will help you think about the data

## II. Make a Picture

it may show unexpected features

## III. Make a Picture

it will help you tell others what you've found.

**These are made easier with technology!**



# Know the Data's W's

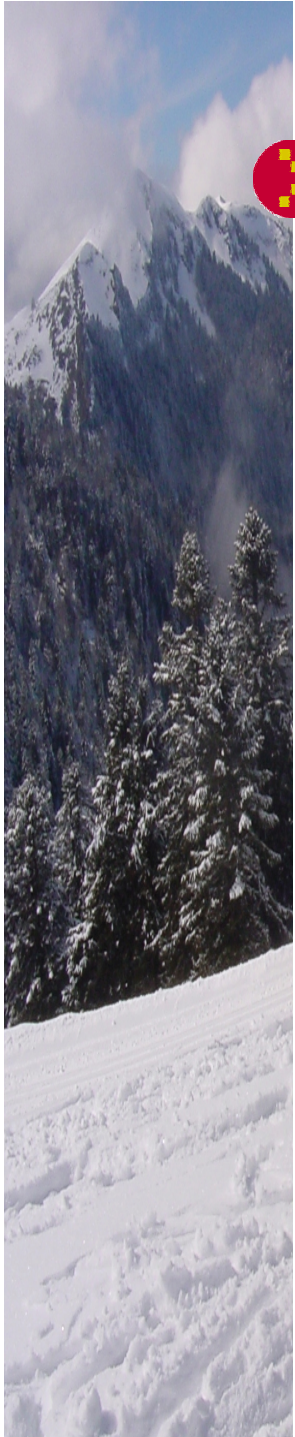
- **Who is the data about?**
  - What's a "row"?
- **What is measured?**
  - What are the "columns"?
  - And in what units?
- **When was it measured?**
- **Where was it measured?**
- **Ho(W) was it measured?**
- **Why was it measured?**



# The W's



Year	Winner	Country	Time	Speed	Stages	Dis (km)	Start	Finish
1903	<b>Maurice Garin</b>	France	94.33.00	25.3	6	2428	60	21
1904	<b>Henri Cornet</b>	France	96.05.00	24.3	6	2388	88	23
1905	<b>Louis Trousselier</b>	France	112.18.09	27.3	11	2975	60	24
1906	<b>Rene Pottier</b>	France	185.47.26	24.5	13	4637	82	14
1907	<b>Lucien Petit-Breton</b>	France	156.22.30	28.5	14	4488	93	33
1908	<b>Lucien Petit-Breton</b>	France	156.09.31	28.7	14	4488	114	36
...								
...								
1999	<b>Lance Armstrong</b>	USA	91.32.16	40.3	20	3687	180	141
2000	<b>Lance Armstrong</b>	USA	92.33.08	39.56	21	3662	180	128
2001	<b>Lance Armstrong</b>	USA	86.17.28	40.02	20	3453	189	144
2002	<b>Lance Armstrong</b>	USA	82.05.12	39.93	20	3278	189	153
2003	<b>Lance Armstrong</b>	USA	83.41.12	40.94	20	3427	189	147
2004	<b>Lance Armstrong</b>	USA	83.36.02	40.53	20	3391	188	147



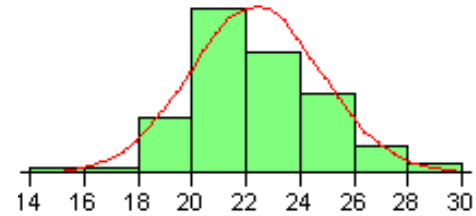
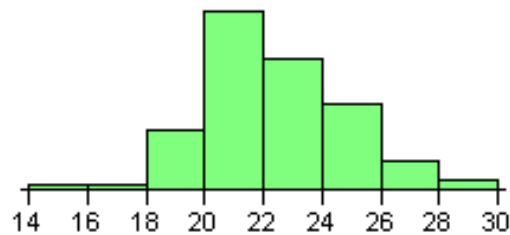
# The Model

- **Statistics is about models**
- **A model is a simplification of reality.**
- **We know it's not perfect**
- **Two quotations from George Box**

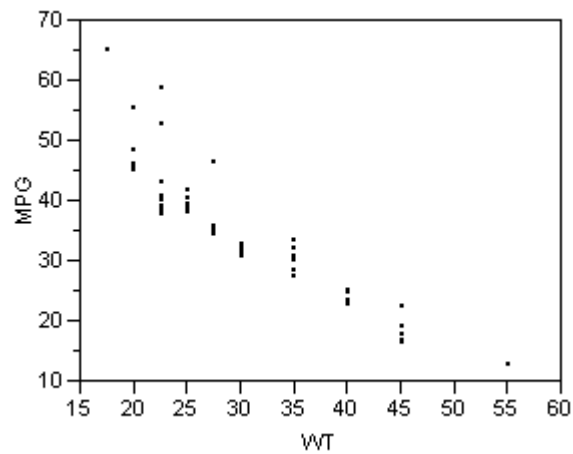


# Common Models

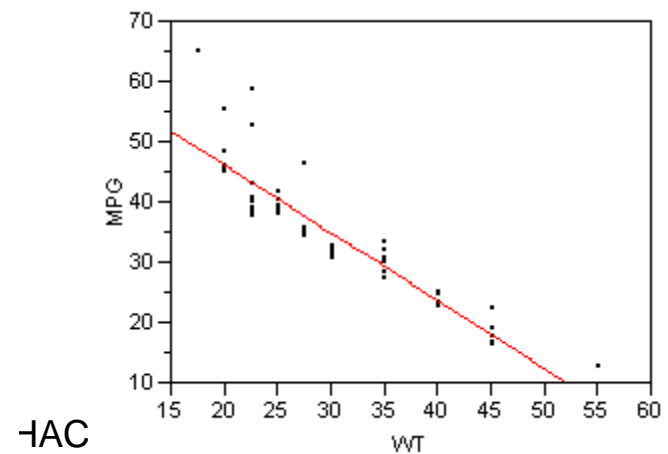
- Probability models



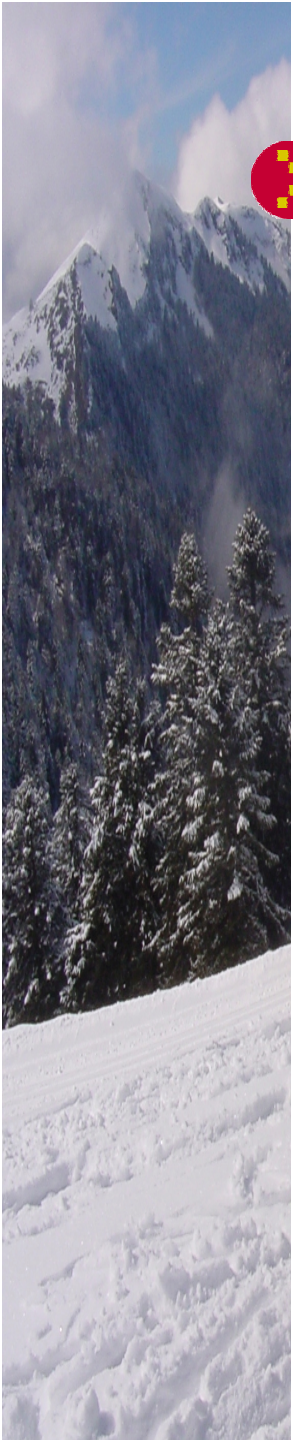
- Regression model



September '1

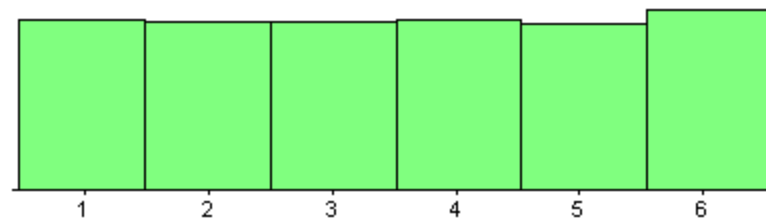
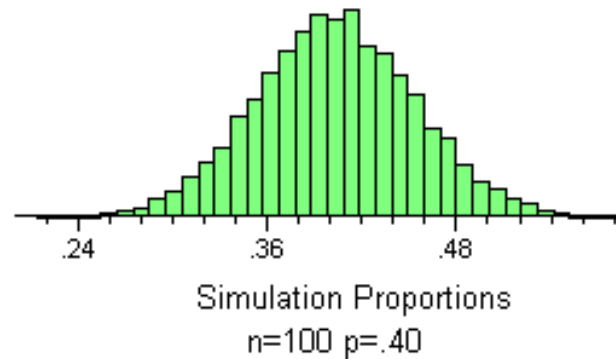


-IAC



# Common Models

- Simulation





# “Pay Dirt” Models

- **Sampling distribution models**
  - By now students know that models are idealized
  - They’ve seen probability models and simulations: CLT follows naturally
- **Null hypothesis models**
  - Wrong (we hope) but useful



# Models...

## Require assumptions

Because they are idealized, they are only really true under idealized assumptions

## Are described by parameters

Parameters refer to models of populations, not to the populations themselves



# Assumptions and Conditions

- **Some assumptions we must just assume. (Pretend)**
- **Many can be checked for plausibility with appropriate conditions**
  - Often the conditions are graphical (Remember the 3 rules)
- **Few are really true**



# Conditions to Check

- **Summary statistics**
  - Quantitative data condition.

Variable -- TCODE	
Mean	54.41
Std Dev	957.50
Std Err Mean	3.11
upper 95% Mean	60.51
lower 95% Mean	48.31

- **T-test**
  - Assumption is that data are Normal
    - Rule of thumb? 30? 50? 100?
    - Nearly normal condition--make a picture



# Show

September 18, 2004

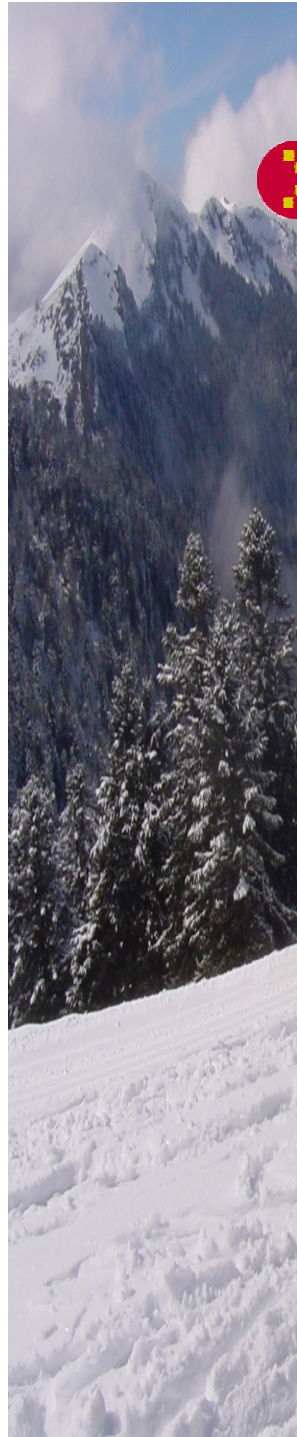
AWL Workshop -- HACC

43



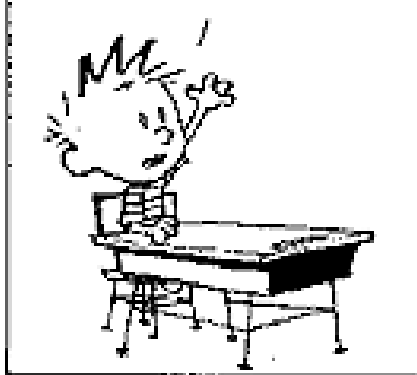
# Show, with Technology

- **Calculation is for calculators and statistics packages.**
  - Let them do it, so students can think about statistical thinking.
  - Show generic output rather than a particular package.
  - Let them do it so we can “play Statistics”



# Play Stats

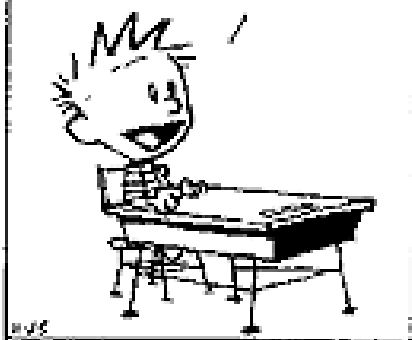
MISS WOODWOOD. MY DAD SAYS WHEN HE WAS IN SCHOOL, THEY TAUGHT HIM TO DO MATH ON A SLIDE RULE.



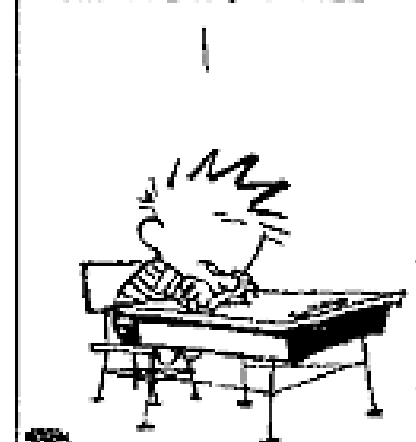
HE SAYS HE HASN'T USED A SLIDE RULE SINCE, BECAUSE HE GOT A FIVE-BUCK CALCULATOR THAT CAN DO MORE FUNCTIONS THAN HE COULD FIGURE OUT IF HIS LIFE DEPENDED ON IT.



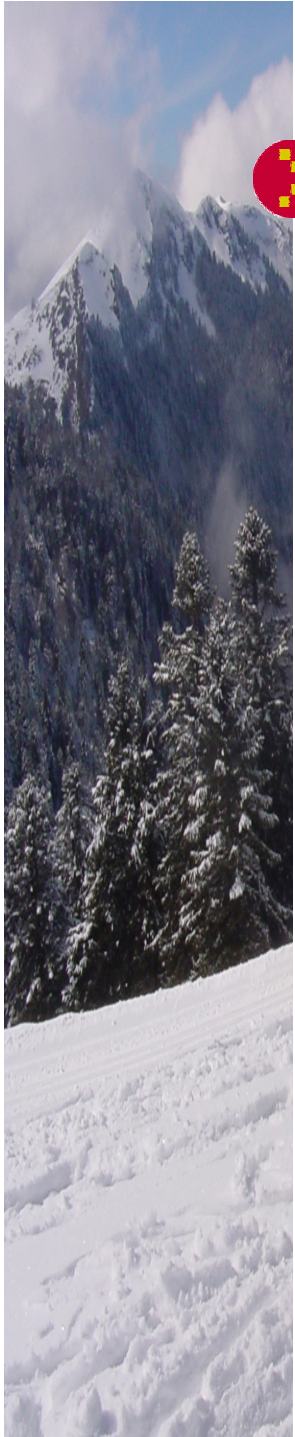
GIVEN THE PACE OF TECHNOLOGY, I PROPOSE WE LEAVE MATH TO THE MACHINES AND GO PLAY OUTSIDE.



MY BILLS ALWAYS DIE IN SUBCOMMITTEE.

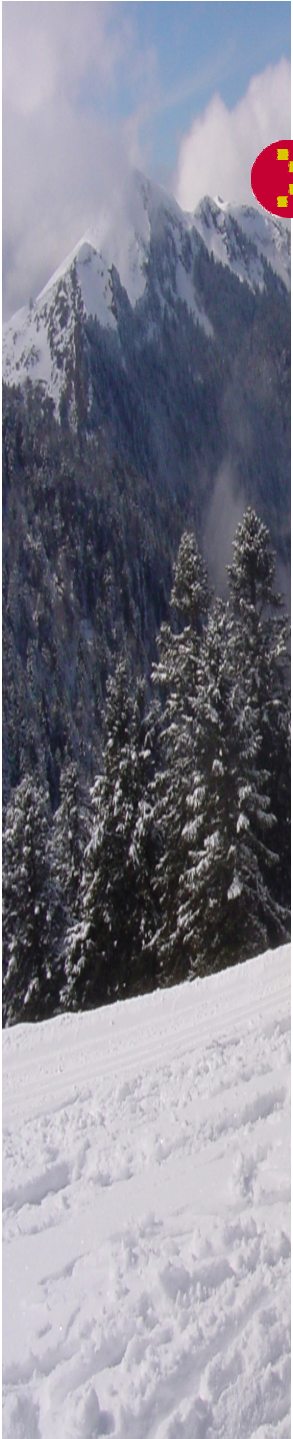


September 18, 2004



# More Help – Reality Checks

- The answer is wrong if it makes no sense -- even if you pushed the buttons you meant to push or gave the command you intended
- Check that the results are plausible
  9. **Professors.** A friend tells you about a recent study dealing with the number of years of teaching experience among current college professors. He remembers the mean but can't recall whether the standard deviation was 6 months, 6 years, or 16 years. Tell him which one it must have been, and why.
- Remember the units!



# Telli

September 18, 2004

AWL Workshop -- HACC

47



# Draw Conclusions

- **Plot the data, but then say what you see.**
  - Give guidance for how to “see”
- **Reject the null hypothesis, but then provide a CI to assess effect size.**
  - Emphasize interplay between tests and CI
- **Think about costs and consequences.**
  - Don’t be satisfied with “I rejected  $H_0$ ”



# What Can Go Wrong?

- **Acknowledge common misapplications and misinterpretations of statistics.**
- **(Hope to) Minimize them in Telling what was found.**



## What Can Go Wrong?

There are many ways in which data that appear at first to be good candidates for regression analysis may be unsuitable. And there are ways that people use regression that can lead them astray. Here's an overview of the most common problems. We'll discuss these at length in the next chapter.

- *Don't fit a straight line to a nonlinear relationship.* Linear regression is suited only to relationships that are, well, *linear*. Fortunately, we can often improve the linearity easily by using re-expression. We'll come back to this topic in Chapter 10.
- *Beware extraordinary points.* Data values can be extraordinary in a regression in two ways. They can have  $y$ -values that stand off from the linear pattern suggested by the bulk of the data, or extreme  $x$ -values. Both kinds of extraordinary points require attention.
- *Don't extrapolate beyond the data.* A linear model will often do a reasonable job of summarizing a relationship in the narrow range of observed  $x$ -values. Once we have a working model for the relationship, it's tempting to use it. But beware of predicting  $y$ -values for  $x$ -values that lie outside the range of the original data. The model may no longer hold there, so such **extrapolations** too far from the data are dangerous.



The  $R^2$  does **not** mean that *protein* accounts for 69% of the *fat* in a BK food item. It is the *variation* in fat content that is accounted for by the linear model.

- *Don't infer that  $x$  causes  $y$  just because there is a good linear model for their relationship.* We have seen that when two variables are strongly correlated, it is often tempting to assume a causal relationship between them. Putting a regression line on a scatterplot tempts us even further, but it doesn't make the assumption of causation any more valid.

Although  $R^2$  measures the *strength* of the linear association, a high  $R^2$  does not demonstrate the *appropriateness* of the regression. A single outlier, or data that separate into two groups rather than a single cloud of points, can make the  $R^2$  seem quite large when, in fact, the linear regression model is simply inappropriate. Conversely, a low  $R^2$  value may be due to a single outlier as well. It may be that most of the data fall roughly along a straight line with the exception of a single point.



# Step-By-Step

- Encourage students to bring all of these ideas together when they solve a statistical problem.
- Illustrate how, step-by-step

## Regression **STEP-BY-STEP**

Even if you hit the fast food joints for lunch, you should have a good breakfast. Researchers recorded facts about 77 breakfast cereals, including the *calories* and *sugar* content (in grams) of a serving. Let's build a linear model to understand how calories are related to sugar content.

**Think**

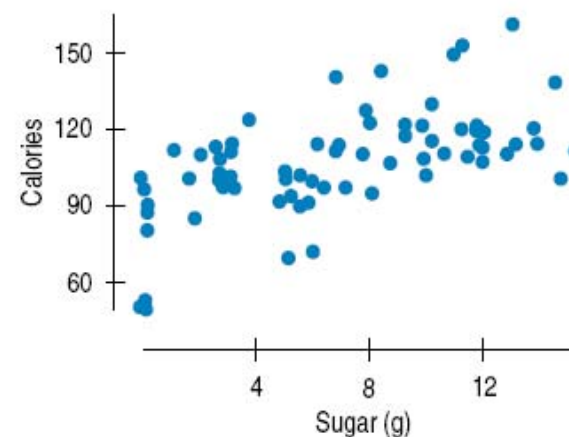
**Variables** Name the variables, report the W's, and specify the questions of interest.

**Plan** To check the conditions for a regression, always make a picture. Never fit a regression without looking at the scatterplot first.

Let  $x = \text{sugar content of cereals (grams)}$   
 $y = \text{calories}$

We have two quantitative variables measured on 77 breakfast cereals.

We are interested in the relationship between these two variables.



We can see from the scatterplot that the direction of the relationship is positive and the form of the association is straight enough.

There are no obvious outliers or groups.

Because the **straight enough condition** is satisfied, we can fit a regression model to these data.

## Show

**Mechanics** If there are no clear violations of the condition, fit a straight line model of the form  $\hat{y} = b_0 + b_1x$  to the data. Summary statistics give the building blocks of the calculation.

Find the slope.

Find the intercept.

Write the equation using meaningful variable names.

State the value of  $R^2$ .

### Calories

$$\bar{y} = 107.0 \text{ calories}$$

$$s_y = 19.5 \text{ calories}$$

### Sugars

$$\bar{x} = 7.0 \text{ grams}$$

$$s_x = 4.4 \text{ grams}$$

### Correlation

$$r = 0.564$$

$$b_1 = \frac{rs_y}{s_x} = \frac{0.564(19.5)}{4.4}$$

$$= 2.50 \text{ calories per gram of sugar.}$$

$$b_0 = \bar{y} - b_1\bar{x} = 107 - 2.50(7) = 89.5 \text{ calories.}$$

So the least squares line is

$$\hat{y} = 89.5 + 2.50x,$$

$$\text{or } \widehat{\text{calories}} = 89.5 + 2.50 \text{ sugar.}$$

Squaring the correlation gives

$$R^2 = 0.564^2 = 0.318 \text{ or } 31.8\%.$$

## Tell

**Interpretation** Describe what the model says in words and numbers. Be sure to use the names of the variables and their units.

The key to interpreting a regression model is to start with the phrase “ $b_1$   $y$ -units per  $x$ -unit,” substituting the estimated value of the slope for  $b_1$  and the names of the respective units. The intercept is then a starting or base value.

$R^2$  gives the fraction of the variability of  $y$  accounted for by the linear regression model.

## Think Again

**Check Again** Even though we looked at the scatterplot *before* fitting a regression model, a plot of the residuals is an essential part of any regression analysis because it is the best check for additional patterns and interesting quirks in the data.

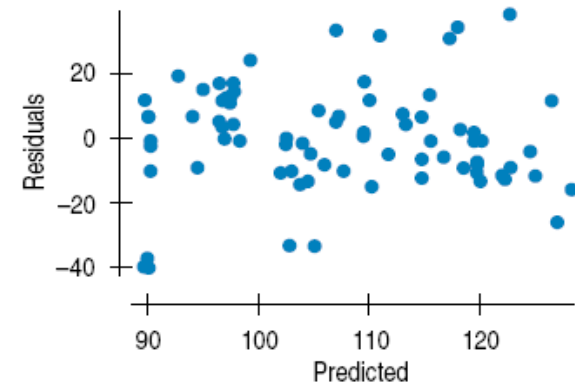
The scatterplot shows a positive, linear relationship and no outliers. The least squares regression line fit through these data has the equation

$$\widehat{\text{calories}} = 89.5 + 2.50 \text{ sugar}.$$

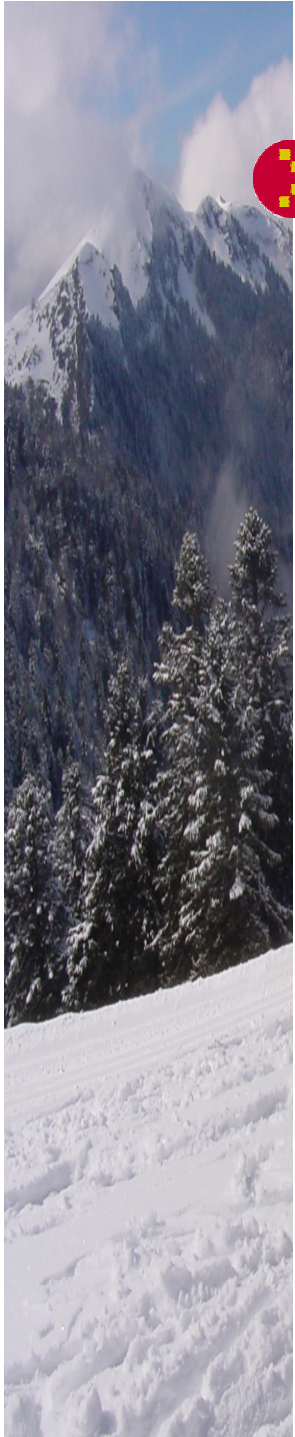
The slope says that cereals gain about 2.50 calories per gram of sugar.

The intercept predicts that sugar-free cereals would average about 89.5 calories.

The  $R^2$  says that 31.8% of the variability in calories is accounted for by variation in sugar content.

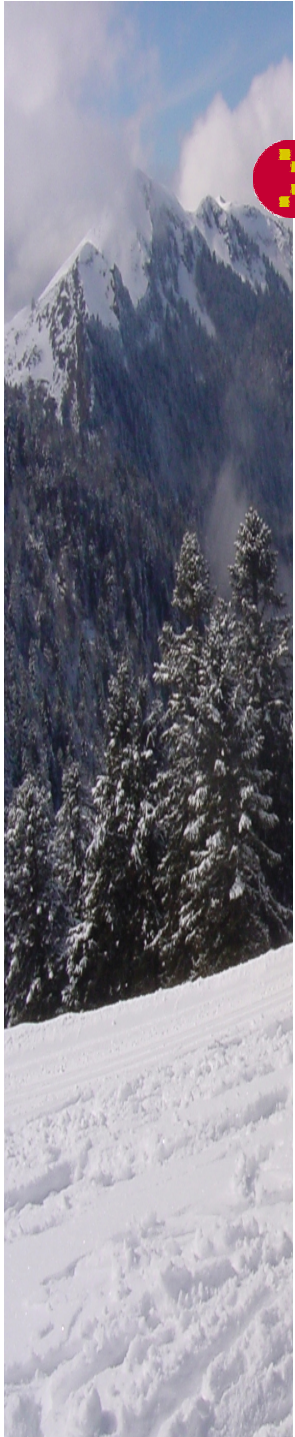


The residuals show a horizontal direction, a shapeless form, and roughly equal scatter for all predicted values. Our linear model appears to be appropriate.



# Take Home Messages

- **Stats is about the real world:**
  - **Technology frees the student to think about the world**
  - **Give the student a structure for a chaotic world**
  - **Root the course in examples taken from the students' lives to make the connection apparent**
  - **Help them with unnatural thinking**



# Thank you !!

September 18, 2004

AWL Workshop -- HACC

56