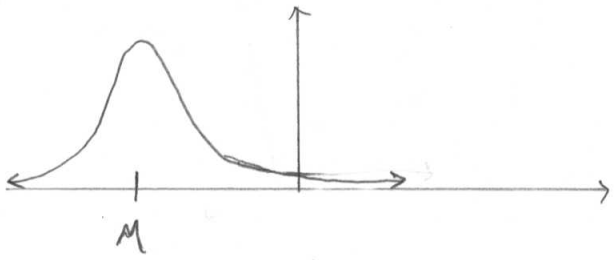
1/27/15 263 Lecture 4

Announcements:
- Chapter 2 WebAssign DUE FRI 1/29/14
- Excel HW 2 Posted, DUE NEXT TUE (2/3) Topic is Scatter Plots & regressions
- EXAM 1 is 1 week from Thursday
- Tomorrow (1/28) is the last day to DROP on VAccess w/out a "W".

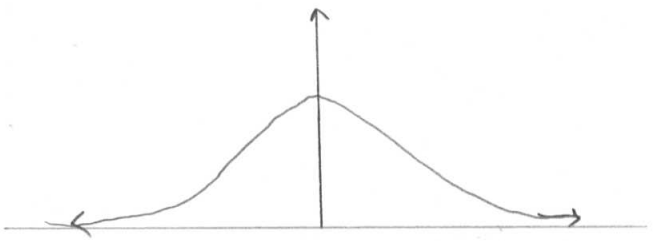
Last Time

$N(\mu, \sigma)$:



"Normal Distribution with mean μ, Std dev σ"

$N(0,1)$:



"Standard Normal"

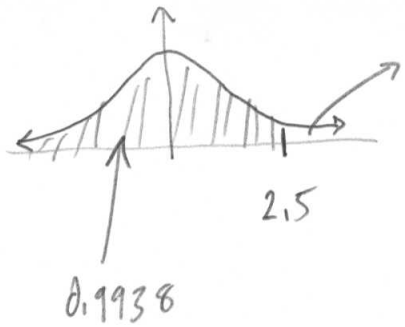
$z = \frac{x-\mu}{\sigma}$ * to Find proportion of a Normally distributed Variable, convert to z-scores & use z-table

* z table gives proportion of z scores <u>to the left</u> of the given z-score.

Ex Suppose $X \sim N(10,2)$. Find % of observations bigger than 15. $z = \frac{15-10}{2} = \frac{5}{2} = 2.5$.

Go to z table:

	0.0
...	
2.5	0.9938



Above $z = 2.5$ is $1 - 0.9938 = 0.0062$

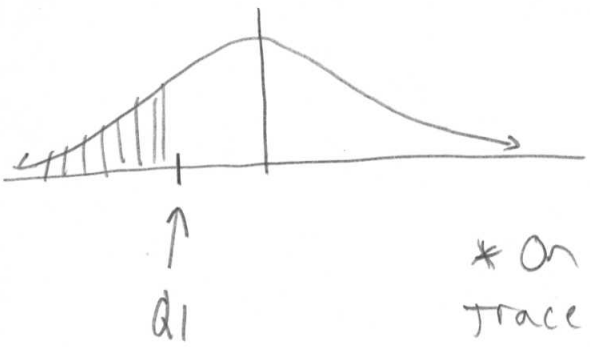
So **0.62%**

On Calculator:

2ND: VARS (DISTR) → normal cdf(a, b, μ, σ)

will give area between a & b under $N(\mu, \sigma)$.

Inverse: What is the IQR of $N(0,1)$?



Q1 is where "CDF" (cumulative area) is 0.25

* On z-Table: look for # in table, trace back to row & col.

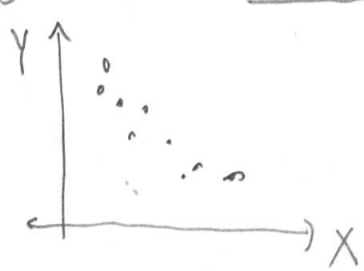
* On Calc: DISTR: invNorm(0.25, 0, 1)

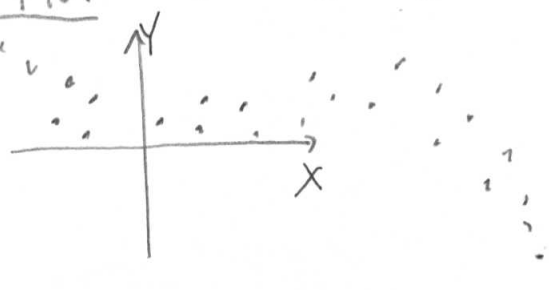
Area μ σ

Chapter 2: Relationships

Two variables can be Associated

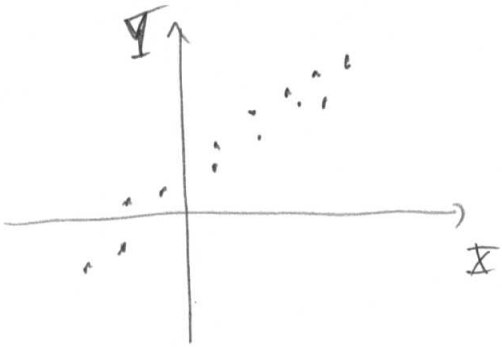
IE Their Scatter Plot shows some Pattern:





Associations can be weak:

Strong:



Linear:

Nonlinear:

For linear associations, we can have either a positive or negative association

"An increase in the cost of gas will decrease the total miles driven"

"An increase in the # of calories consumed will increase your BMI"

Measuring amount of linear association:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} z_{y_i}$$

$$z_{x_i} = \frac{x_i - \bar{x}}{s_x}$$

$$z_{y_i} = \frac{y_i - \bar{y}}{s_y}$$

- If $z_{x_i} z_{y_i} > 0$, There is a Positive association between X_i & Y_i (1st or 3rd Quad on re-centered Scatter Plot)

- If $z_{x_i} z_{y_i} < 0$, There is a negative association between X_i & Y_i (2nd or 4th Quad. on Scatter Plot)

r is The Average of all These, So

If "most" $z_{x_i} z_{y_i}$'s Are > 0, $r > 0$

If "most" $z_{x_i} z_{y_i}$'s are < 0, $r < 0$

* $-1 \leq r \leq 1$

* $r = -1$ is Perfect negative Correlation

So $\hat{Y} = mX + b$ w/ $m < 0$

* $r = 1$ is Perfect Positive Correlation

$\hat{Y} = mX + b$ w/ $m > 0$

* $r \approx 0$ no Correlation

- Most of the time, $-1 < r < -0.5$ or $0.5 < r < 1$, which means

That a linear model might be a good fit

IE $Y \simeq mX + b$ is a good guess

" If I increase X, Y will increase (decrease) by about $m\Delta X$ "

↑ change in X.

Notes about r:

- only measures linear association
- Both X & Y should be Quantitative
- r is not robust

Linear Regression

- We Have a data set with 2 variables X & Y, And we Think Y might be adequately described by $\boxed{Y = mX + b}$ (really, We Think $Y \approx mX + b$ with small error!)

- m is The slope - it is interpreted as

"When X increases by 1 unit, Y increases or decreases by m units"

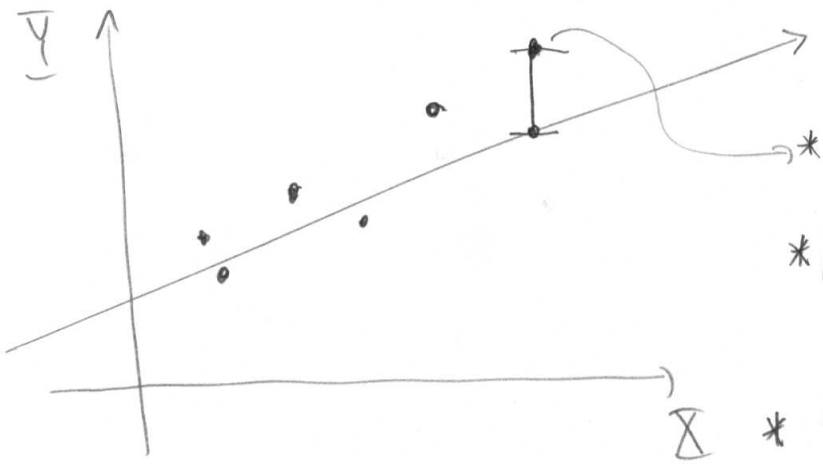
IE $F = 1.8C + 32$ ($F \sim$ °Farenheit. $C \sim$ °Celcius)

- We would Proceed w/ a linear regression if we had a Correlation Close to ± 1

- How do we find the "Best fit line"?

- Want to "minimize distance from line to data set"

- Standard (aka old) method: least squares



$\underline{Y} = m\underline{X} + b$

* Data: (x_i, y_i)

* linear model: $(x_i, mx_i + b)$

* squared error: $(y_i - (mx_i + b))^2$

least squares minimizes $\sum \text{error}^2$

$\hat{\underline{Y}} = m\underline{X} + b$ where $m = r\frac{S_Y}{S_X}$

$b = \bar{y} - m\bar{x}$

* Book uses $y = b_1 X + b_0$

$r \sim$ Correlation Coeff

$S_X, S_Y \sim$ Std deviations of X & Y

$\bar{X}, \bar{Y} \sim$ means of X, Y

How Can We Use a least Squares lines?

* Interpolation (Don't have data point for $X=7$, what should it be?)
* Not Extrapolation

Example using The Fruits/veggies vs Smoking data, we have

$\bar{X} = 23.9$ $\quad S_X = 4.37$

$\bar{Y} = 15.33$ $\quad S_Y = 3.58$

$r = -0.745$

So, $Y = mX + b$ w/ $m = r\frac{S_Y}{S_X} = -0.745\frac{3.58}{4.37} \simeq -0.61$

$b = 15.33 - (-0.61)23.9$

$= 29.91$

So $\boxed{Y = -0.61X + 29.91}$ (to 2 Dec Places)

Interpretation:

If a State increases it's Average fruit & veggie In take by 1 Per day, 0.61% Fewer People will Smoke everyday (causation??)

Facts About linear regression

(1) Alternate Interpretation of Slope:

$m = r\frac{S_y}{S_x}$ If X increases by 1 std dev., Y will increase (decrease) by r std deviations

IE

Student	X midterm	Y final
1	65	77
2	43	52
3	86	84
...		

$\bar{X} = 70 \quad S_x = 10$

$\bar{Y} = 72 \quad S_y = 8$

$r = 0.5$

If a student's midterm score was 10 pts higher, we would expect their final exam score to be about 4 points higher.

(2) How good is the fit?

r^2 is the fraction of variation in Y that is explained by the least-squares fit.

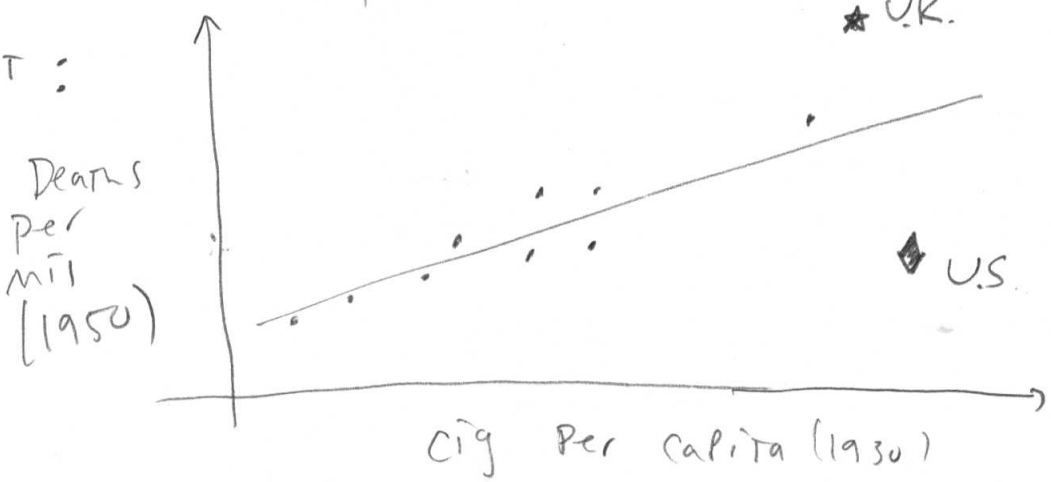
Eg. For fruits/smoking, $r^2 = 0.55$. So about 55% of the variation in Y is explained by variation in X, via the least sq. line.

(3) LS line always passes through $(\bar{X}, \bar{Y})$

(4) OUTLIERS: Consider the following data.

	Cig Per Capita (1930)	Lung Cancer death Per Million (1950)
Iceland	220	58
Norway	250	90
Sweden	310	115
Denmark	380	165
Australia	455	170
Holland	460	245
Canada	510	150
Switzerland	530	250
Finland	1115	350
UK	1145	465
US	1280	190

- SCATTER PLOT:



$Y = 0.22911X + 65.749$

$R^2 = 0.549$

Take out U.S.: $Y = 0.3577X + 13.553$

$R^2 = 0.8855$

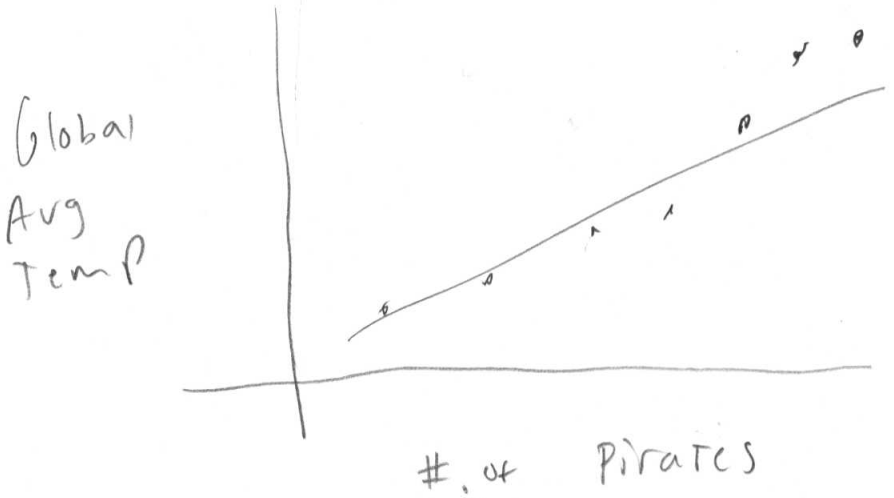
Warnings about Regression / Causation:

- "Everything" is a line if you get "localized" enough! ("Restricted range")
- Inspect Residuals to see Patterns:

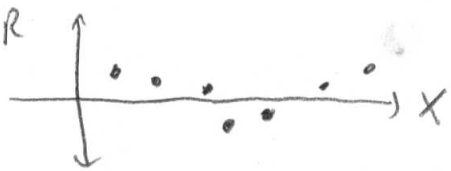
  Residual = Observed y - Predicted $\hat{y}$

- Lurking Vars

  Pirates Cause Global Warming:



EX (Residuals) Sketch a graph of the scatter Plot given the Residuals:



Section 2.7: Read on your own. Correlation Is Necessary For causation, but NOT sufficient. Experiment OR-5 criteria: 1) Strong Assoc. 2) Consistent across many studies 3) Strong response 4) Time 5) Plausibility