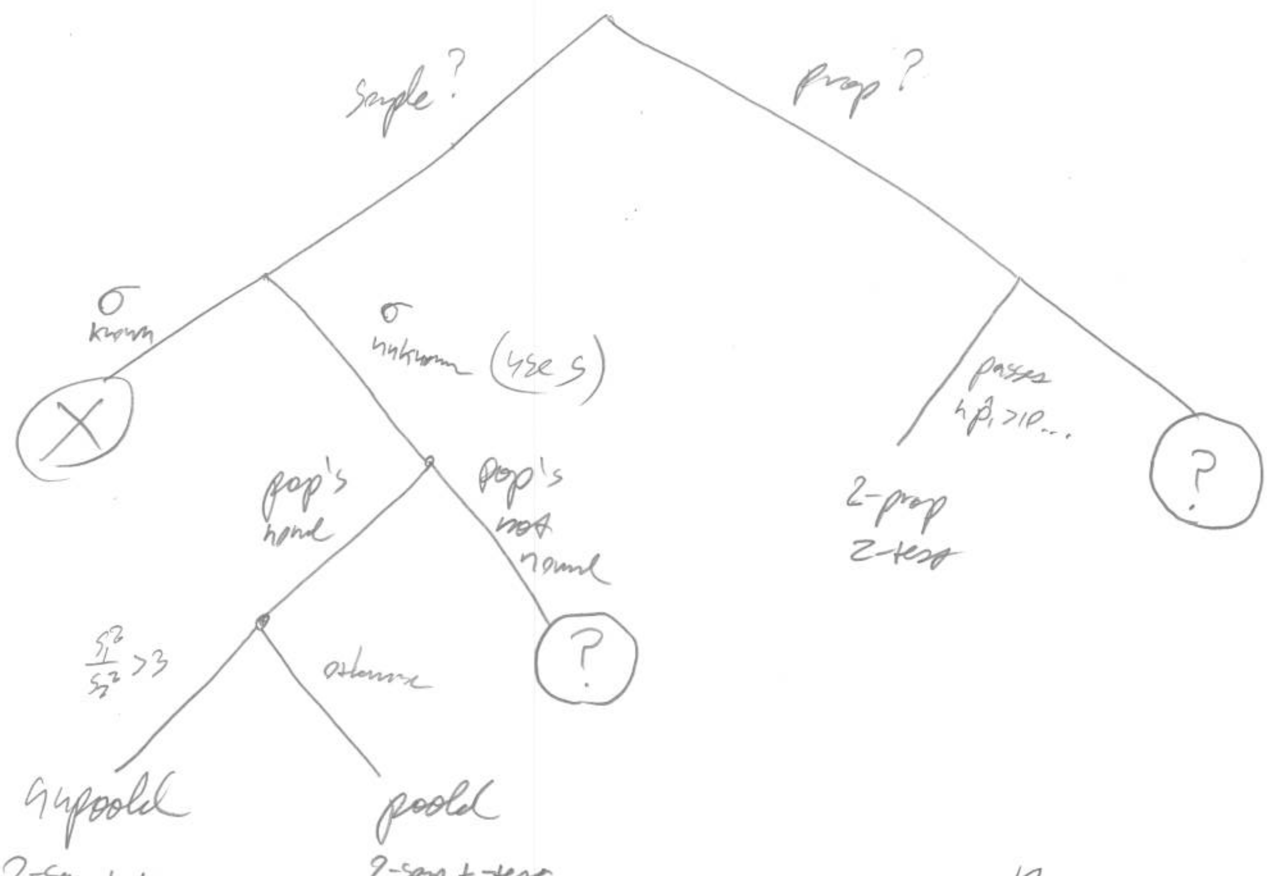
6/28/11

Lecture #18

Plan

Tree for ~~2~~ bivariate hyp. testing



Move to p2

We've seen X, Y ind means by def that

$$P(X=x, Y=y) = P(X=x)\,P(Y=y)$$ for all x, y in the support of X, Y

joint prob. distr. How to disprove? $X, Y \sim \text{Bernoulli}(\frac{1}{2})$

$S = X + Y$ Prove S, X not independent

Yesterday, we proved $E[XY] = E[X]E[Y]$

let's see this in an example:

$$X, Y \overset{iid}{\sim} \begin{cases} 0 & \text{w.p. } 1/2 \\ 1 & \text{w.p. } 1/3 \\ 2 & \text{w.p. } 1/6 \end{cases}$$

$E[X] = E[Y]$
$= 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{6}$
$= \frac{1}{3} + \frac{1}{3} = \boxed{\frac{2}{3}}$

let $Z = XY$

Z

Y \ X	0	1	2
0	0	0	0
1	0	1	2
2	0	2	4

Y \ X	0	1	2
0	1/4	1/6	1/12
1	1/6	1/9	1/18
2	1/12	1/18	1/36

$P(X=x, Y=y)$
at the same time
(102)
Joint prob distribution
seen this before

$E[Z] = E[XY] = 1 \cdot \frac{1}{9} + 2 \cdot \frac{1}{18} + 2 \cdot \frac{1}{18} + 4 \cdot \frac{1}{36}$

$= \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} = \frac{4}{9} = E[X]E[Y] = \frac{4}{9}$ ✓

Can we think of model: let $N \sim \text{Poisson}(\lambda)$ (# people / d), $R \sim N(10, 5)$ (revenue per person)
$E[NR]$ = expected revenue per day

Go back to p1 now... So S, X not independent.
Sort of like smoking & lung cancer. How do we measure
dependency between r.v.'s?

ch 10.4 Covariance: expected value of product deviation from means:

$$Cov[X,Y] = E\left[(X-\mu_X)(Y-\mu_Y)\right]$$ a measure of dependence

Cov can be pos or neg unlike variance. If X, Y both weight their stakes above μ's more heavily than below μ then cov will be positive. If X weights above μ and Y weights below μ $\Rightarrow$ Cov is negative.

If X, Y independent, cov is zero:

good alternate formula

$$E\left[(X-\mu_X)(Y-\mu_Y)\right] = E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X\mu_Y =$$
$$= E[XY] - \mu_X\mu_Y - \cancel{\mu_Y\mu_X + \mu_X\mu_Y}$$
$$= E[X]E[Y] - \mu_X\mu_Y \quad \text{if IND}$$
$$= \mu_X\mu_Y - \mu_X\mu_Y = 0 \checkmark$$

X, Y can be high variance but $Var[X+Y] = 0$

for instance let X be a "put" option (buy), Y be a put option (sell).

$$X \sim \begin{cases} 0 & \text{w.p. } \frac{1}{3} \\ 1 & \text{w.p. } \frac{1}{3} \\ 2 & \text{w.p. } \frac{1}{3} \end{cases}$$

$$Y \sim \begin{cases} 0 & \text{w.p. } \frac{1}{3} \\ -1 & \text{w.p. } \frac{1}{3} \\ -2 & \text{w.p. } \frac{1}{3} \end{cases}$$

$X+Y=0$ always $\Rightarrow$ $\mathrm{Var}[X+Y]=0$

$\mathrm{Var}[X] = \frac{1}{3}\left((0-1)^2+(1-1)^2+(2-1)^2\right) = \frac{2}{3}$ $\quad \mu_X = E[X] = 1$ $\quad$ do these first

$\mathrm{Var}[Y] = \frac{1}{3}\left((0--1)^2+(-1--1)^2+(-2--1)^2\right) = \frac{2}{3}$ $\quad \mu_Y = E[Y] = -1$

$\mathrm{Cov}[X,Y] = E\left[(X-\mu_X)(Y-\mu_Y)\right] = E[XY] - \mu_X\mu_Y = E[XY]+1 = -\frac{2}{3}$

XY

$$\begin{array}{cc|ccc} & & & X & \\ & & 0 & 1 & 2 \\ \hline & 0 & 0 & 0 & 0 \\ Y & -1 & 0 & -1 & -2 \\ & -2 & 0 & -2 & -4 \end{array}$$

$P(X=x, Y=y)$

$$\begin{array}{cc|ccc} & & & X & \\ & & 0 & 1 & 2 \\ \hline & 0 & 1/3 & 0 & 0 \\ Y & -1 & 0 & 1/3 & 0 \\ & -2 & 0 & 0 & 1/3 \end{array}$$

Dependent!

$E[XY] = \frac{1}{3}(0) + \frac{1}{3}(-1) + \frac{1}{3}(-4) = -\left(\frac{1}{3}+\frac{4}{3}\right) = -\frac{5}{3}$

In general $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{Cov}[X,Y]$ $\quad$ General rule for addition of variances p226

$= \frac{2}{3} + \frac{2}{3} + 2\left(-\frac{2}{3}\right) = 0$

Look at units of covariance $\quad E[XY] - \mu_X\mu_Y$ $\quad \$^2$

Units of X times units of Y

Also if X ranges from -\$1 trillion to \$1 trillion and Y is in microns that makes for a strange unit... Also can't compare $\mathrm{Cov}[X,Y]$ to $\mathrm{Cov}[U,V]$

We need some way to standardize the measure of dependence.
It would be nice to be unitless and in a standard range
so we can make apples-apples comparisons.

"rho"

$$\rho \triangleq \mathrm{Corr}[X,Y] = \frac{\mathrm{Cov}(X,Y)}{SD(X)\,SD(Y)}$$ divide cov by the SD's

"correlation"

Standard between $-1, 1$ ... $\rho \in [-1, 1]$ p327

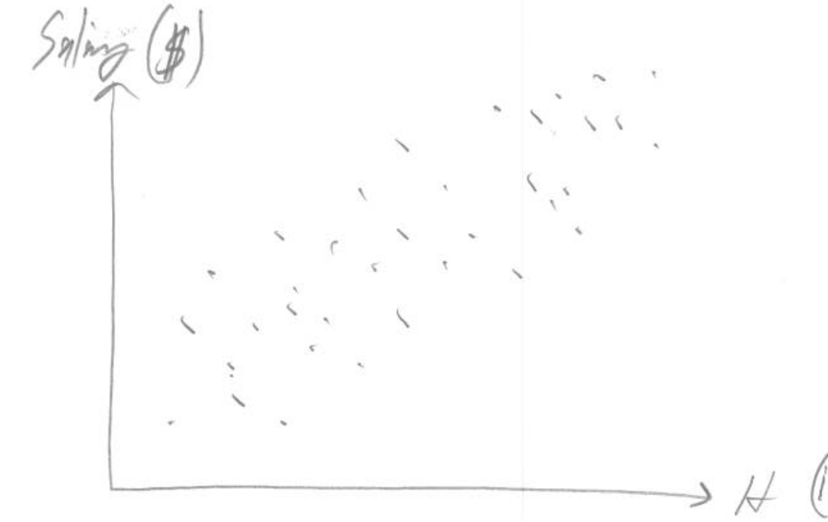
-1 means perfectly negative association as $X\uparrow\ Y\downarrow$ (vice versa), $+1$ means perfectly pos. assoc $X\uparrow, Y\uparrow$ (vice versa)

0 means no correlation at all $X\uparrow\ Y?$ etc

the word is X, Y are uncorrelated.

Why do we care? (Done w/ ch. 10) Bivariate analysis univariate vs. bivariate data

(X, Y) are realized in pairs (x, y). (Ch 6)



Scatterplot

(x_1, y_1)
(x_2, y_2)
$\vdots$
(x_n, y_n)

Sample size n

We have all the std summary ~~stats~~ stats $\bar{x}, \bar{y}, s_x, s_y$
but now we'll have more ... Pearson ...

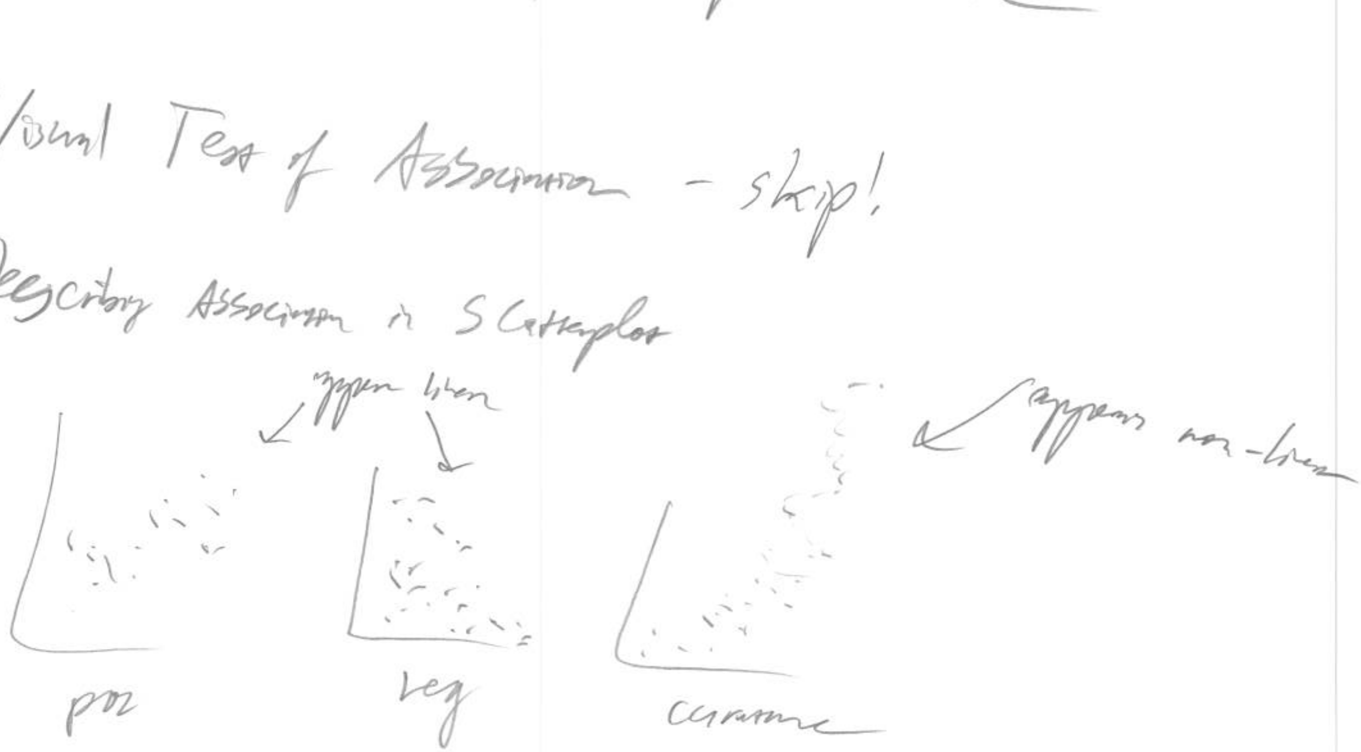
There's a reason I chose x, y axes this way

X: explanatory variable / independent variable

Y: response variable / dependent variable

Visual Test of Association - skip!

Describing Association in Scatterplots



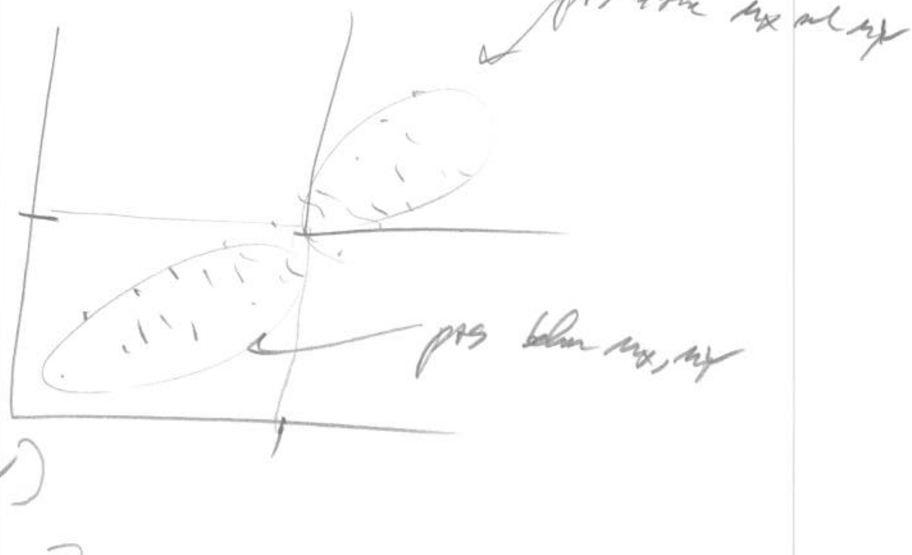
Almost all of Statistical analysis is based on X,Y having a linear association. You'll see this in Stat 102.

both pos, linear, but left has less variance than right

<u>Things to look for</u>

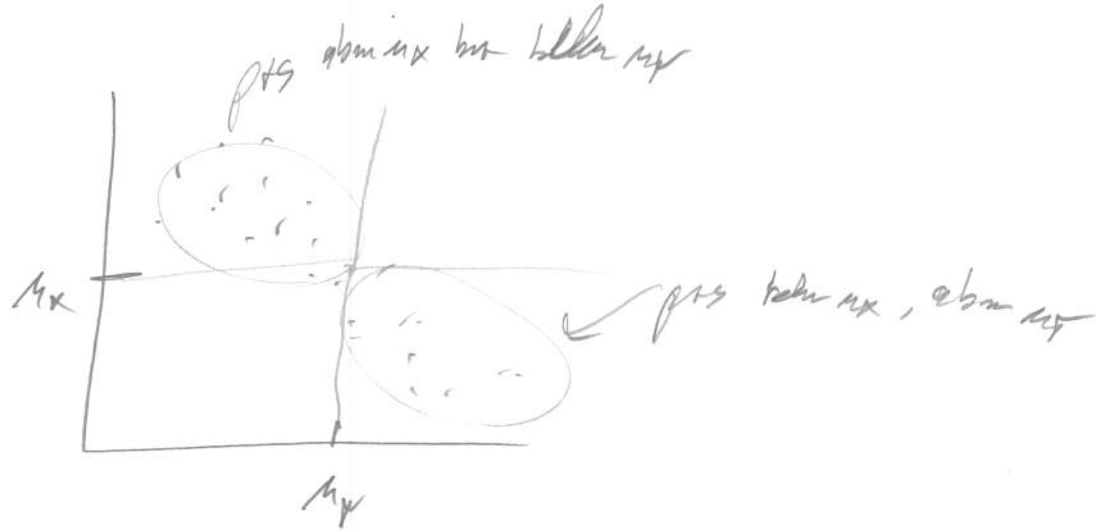
1. Direction
2. Linearity
3. Spread
4. Outliers

Measuring Association



Cov(X,Y)

Think $E\left[(X-m_x)(Y-m_Y)\right] > 0$



$$E\left[(X-\mu_x)(Y-\mu_y)\right] < 0$$

Sample covariance $$S_{XY} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$ p110

→ estimates $Cov[X,Y]$

$$r_{XY} = \frac{S_{XY}}{S_x S_y} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}}$$

→ estimates ρ