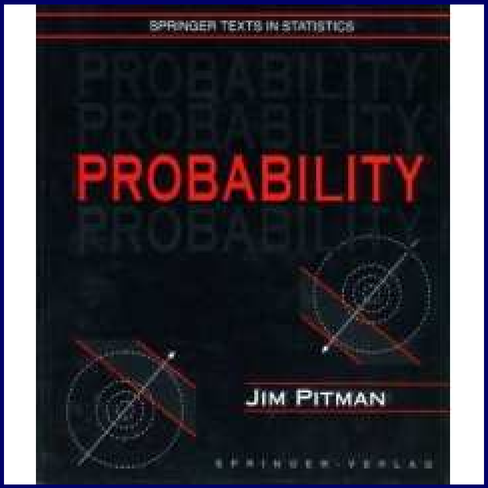
Introduction to probability

Stat 134 FAll 2005

Berkeley



Lectures prepared by:

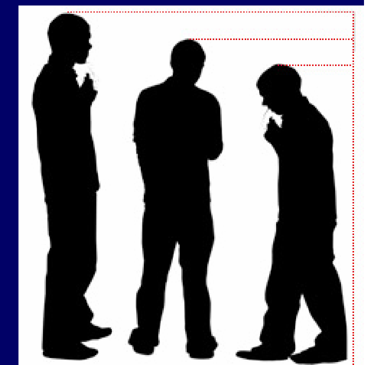
Elchanan Mossel
Yelena Shvets

Follows Jim Pitman's book:

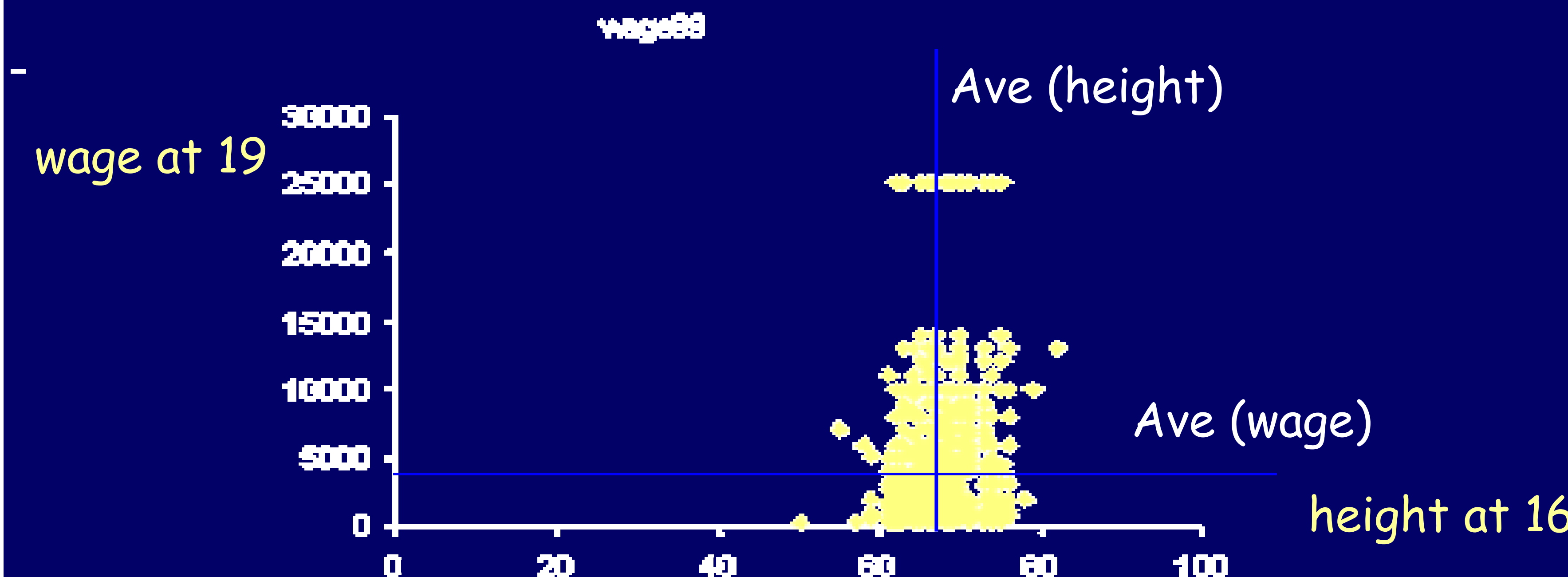
Probability

Sections 6.4

Do taller people make more money?



Question: How can this be measured?



National Longitudinal Survey of Youth 1997 (NLSY97)

Definition of Covariance

$$\text{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_y)]$$

Alternative Formula

$$\text{Cov}(X,Y) = E(XY) - E(X)E(Y)$$

Variance of a Sum

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X,Y)$$

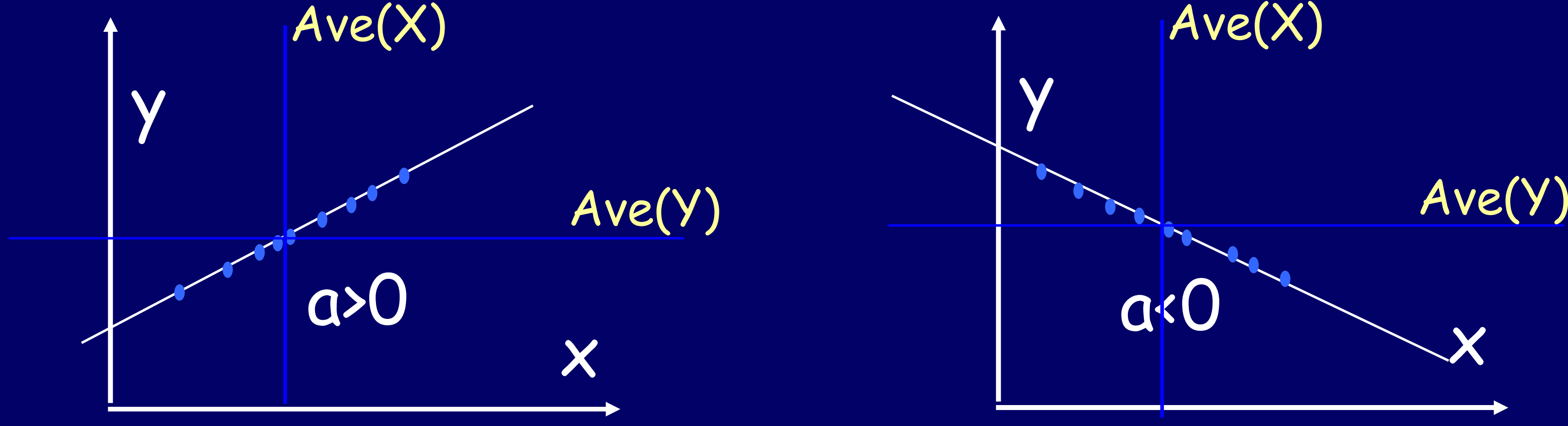
<u>Claim:</u> Covariance is Bilinear

$$\begin{aligned}\text{Cov}(aX + b, cY + d) &= E[(aX - E(aX))(cY - E(cY))] \\ &= E[ac(X - \mu_x)(Y - \mu_y)] \\ &= ac\text{Cov}(X, Y).\end{aligned}$$

What does the sign of covariance mean?

Look at $Y = aX + b$.

Then: $Cov(X,Y) = Cov(X,aX + b) = aVar(X)$.



If $a > 0$, above the average in X goes with above the ave in Y.

If $a < 0$, above the average n X goes with below the ave in Y.

$Cov(X,Y) = 0$ means that there is no linear trend which connects X and Y.

Meaning of the value of Covariance

Back to the National Survey of Youth study :

the actual covariance was 3028 where height is inches and the wages in dollars.

Question: Suppose we measured all the heights in centimeters, instead. There are **2.54** cm/inch?

Question: What will happen to the covariance?

Solution: So let H_I be height in inches and H_C be the height in centimeters, with W – the wages.

$$Cov(H_C,W) = Cov(2.54\ H_I,W) = 2.54\ Cov\ (H_I,W).$$

So the value depends on the units and is not very informative!

Covariance and Correlation

Define the <u>correlation coefficient</u>:

$$\rho = \mathrm{Corr}(X, Y) = E\left(\frac{X - E(X)}{SD(X)} \cdot \frac{Y - E(Y)}{SD(Y)}\right)$$

Using the linearity of Expectation we get:

$$\rho = \frac{\mathrm{Cov}(X, Y)}{SD(X)SD(Y)}$$

Notice that $\rho(aX+b, cY+d) = \rho(X,Y)$. This new quantity is <u>independent of the change in scale</u> and it's value is quite informative.

Covariance and Correlation

Properties of correlation:

$$X^* = \frac{(X - \mu_X)}{SD(X)} \text{ and } Y^* = \frac{(Y - \mu_Y)}{SD(Y)}$$

$$E(X^*) = E(Y^*) = 0 \text{ and } SD(X^*) = SD(Y^*) = 1$$

$$Corr(X, Y) = Cov(X^*, Y^*) = E(X^* Y^*)$$

Covariance and Correlation

Claim: The correlation is always between –1 and +1

$$E(X^{*2}) = E(Y^{*2}) = 1$$

$$0 \le E(X^* - Y^*)^2 = 1 + 1 - 2E(X^*Y^*)$$

$$0 \le E(X^* + Y^*)^2 = 1 + 1 + 2E(X^*Y^*)$$

$$-1 \le E(X^*Y^*) \le 1$$

$$-1 \le \mathrm{Corr}(X, Y) \le 1$$

$$\boxed{\rho = 1 \text{ iff } Y = aX + b.}$$

Correlation and Independence

X & Y are uncorrelated iff any of the following hold

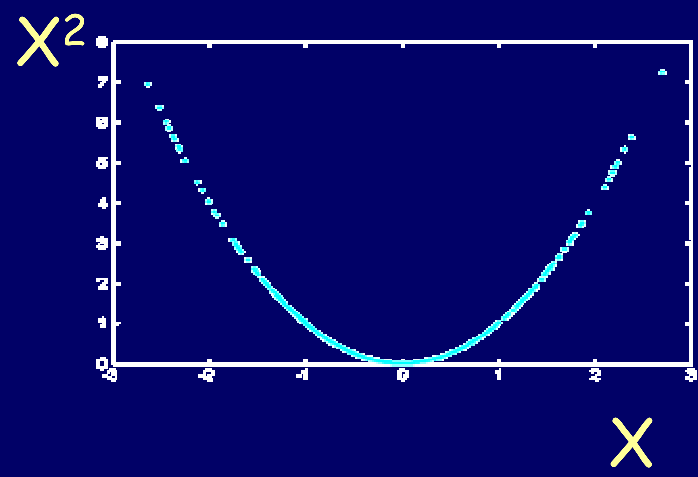
$$Cov(X,Y) = 0,$$

$$Corr(X,Y) = 0$$

$$E(XY) = E(X)\,E(Y).$$

In particular, if X and Y are independent they are uncorrelated.

Example: Let $X \sim N(0,1)$ and $Y = X^2$, then $Cov(XY) = E(XY) - E(X)E(Y) = E(X^3) = 0$, since the density is symmetric.



Roll a die N times. Let X be #1's, Y be #2's.

Question: What is the correlation between X and Y?

Solution:

To compute the correlation directly from the multinomial distribution would be difficult. Let's use a trick:

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y).$$

Since X+Y is just the number of 1's or 2's, $X+Y \sim Binom(p_1+p_2,N)$.

$$Var(X+Y) = (p_1+p_2)(1 - p_1+p_2)\, N.$$

And $X \sim Binom(p_1,N)$, $Y \sim Binom(p_2,N)$, so

$$Var(X) = p_1(1-p_1)N; \quad Var(Y) = p_2(1-p_2)N.$$

Correlations in the Multinomial Distribution

Hence

$$Cov(X,Y) = (Var(X+Y) - Var(X) - Var(Y))/2$$

$$Cov(X,Y) = N((p_1+p_2)(1 - p_1-p_2) - p_1(1-p_1) -p_2(1-p_2))/2 = -N\, p_1\, p_2$$

$$\rho = \frac{-N p_1 p_2}{\sqrt{N p_1 (1 - p_1)}\sqrt{N p_2 (1 - p_2)}}$$

$$= -\sqrt{\frac{p_1 p_2}{(1 - p_1)(1 - p_2)}}$$

In our case $p_1 = p_2 = 1/6$, so $\rho = 1/5$. The formula holds for a general multinomial distribution.

Variance of the Sum of N Variables

$$Var(\sum_i X_i) = \sum_i Var(X_i) + 2 \sum_{j<i} Cov(X_i X_j)$$

Proof:

$$Var(\sum_i X_i) = E[\sum_i X_i - E(\sum_j X_i)]^2$$

$$[\sum_i X_i - E(\sum_j X_i)]^2 = [\sum_i (X_i - \mu_i)]^2$$

$$= \sum_i (X_i - \mu_i)^2 + 2 \sum_{j<i} (X_i - \mu_i)(X_j - \mu_j).$$

Now take expectations and we have the result.

Variance of the Sample Average

Let the population be a list of N numbers x(1), …, x(N). Then

$$\bar{x} = (\textstyle\sum_{i=1}^{N} x(i))/N \qquad \sigma^2 = (\textstyle\sum_{i=1}^{N} (x(i)-\bar{x})^2)/N$$

are the population mean and population variance.

Let $X_1, X_2, \ldots, X_n$ be a sample of size n drawn from this population. Then each X_k has the same distribution as the entire population and

$$E(X_k) = \bar{x} \quad \& \quad Var(X_k) = \sigma^2$$

Let $\bar{X}_n = \dfrac{(X_1+X_2+\ldots+X_n)}{n}$ be the sample average.

Variance of the Sample Average

By linearity of expectation $E(\bar{X}_n) = \bar{x}$, both for a sample drawn <u>with</u> and <u>without replacement</u>.

When $X_1, X_2, \ldots, X_n$ are drawn <u>with replacement</u>, they are independent and each X_k has variance σ^2. Then

$$Var(\bar{X}_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \qquad SD(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

Variance of the Sample Average

Question: What is the SD for sampling without replacement?

$\bar{X}_n = S_n / n$

Solution: Let $S_n = X_1 + X_2 + \ldots + X_n$. Then .

$$Var(S_n) = \sum_i Var(X_i) + 2 \sum_{j<i} Cov(X_i X_j)$$

By symmetry: $Cov(X_i, X_j) = Cov(X_1, X_2)$, so

$$Var(S_n) = n\sigma^2 + n(n-1)\,Cov(X_1 X_2)\,.$$

In particular, $0 = Var(S_N)\ N\sigma^2 + N(n-N)\,Cov(X_1, X_2)$

Therefore: $Cov(X_1 X_2) = -\sigma^2/(N-1)$.

And hence $Var(S_n) = \sigma^2\, n(1 - (n-1)/(N-1))$.

$$Var(\bar{X}_n) = \frac{Var(S_n)}{n^2}; \quad SD(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$