

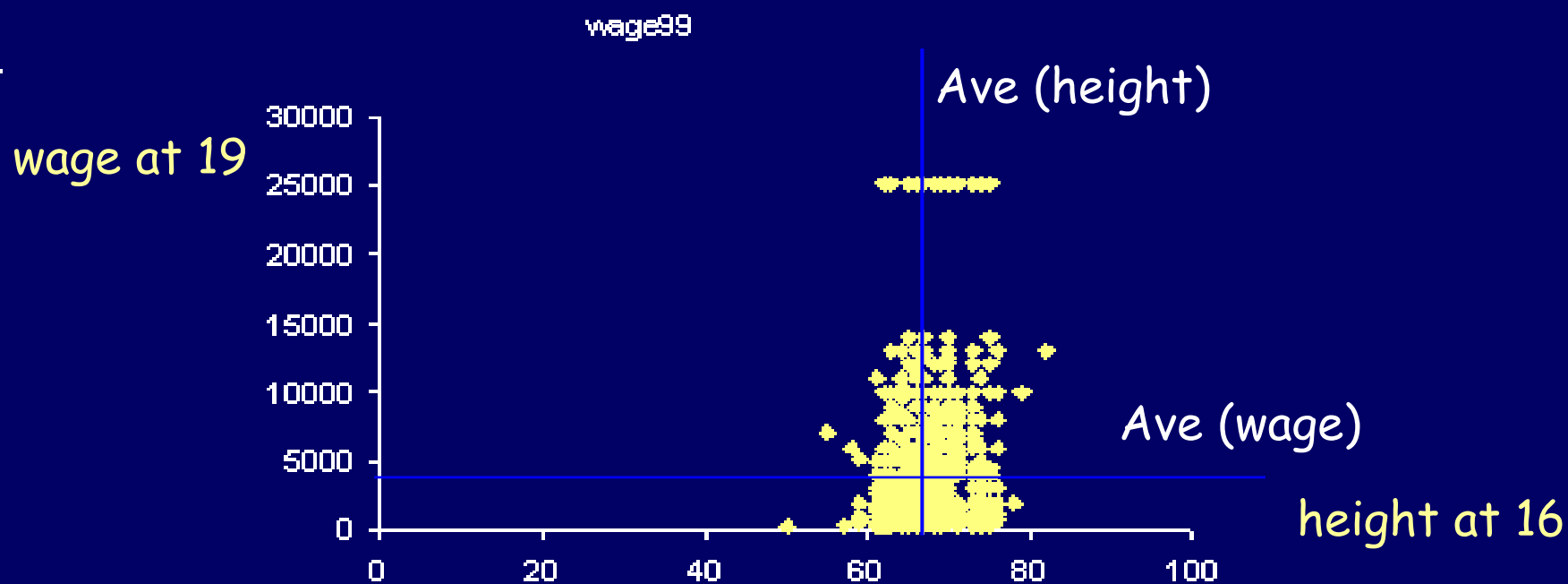
Lectures prepared by:  
Elchanan Mossel  
Yelena Shvets

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

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

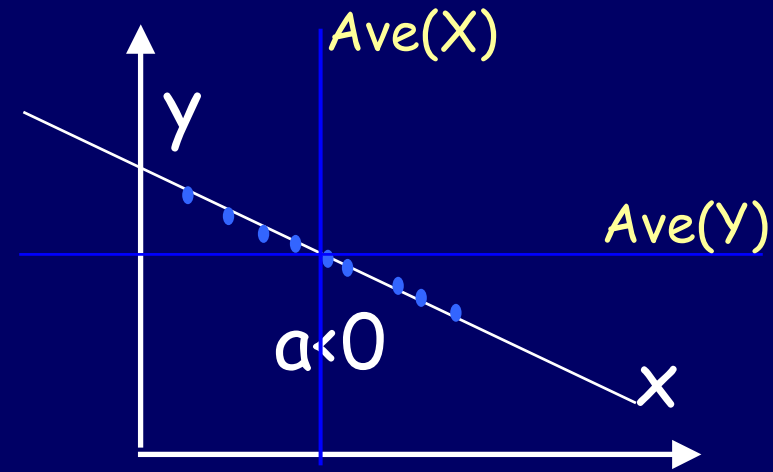
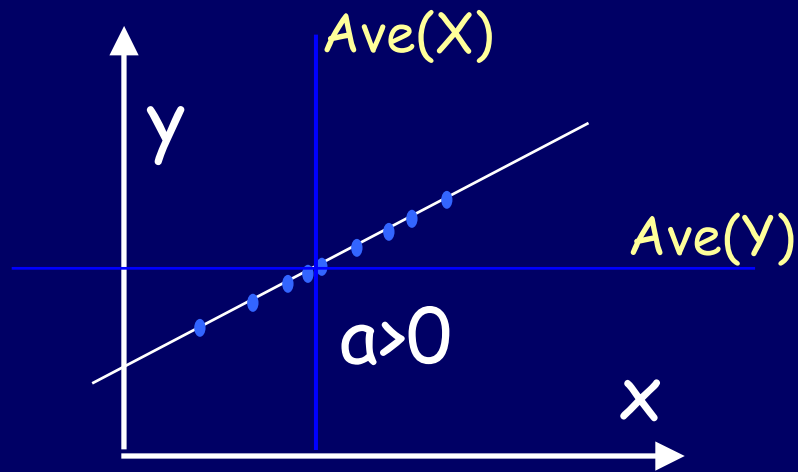
## Claim: Covariance is Bilinear

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

## What does the sign of covariance mean?

Look at  $Y = aX + b$ .

Then:  $\text{Cov}(X, Y) = \text{Cov}(X, aX + b) = a\text{Var}(X)$ .



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

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

$\text{Cov}(X, Y) = 0$  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.

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

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

# Covariance and Correlation

Define the correlation coefficient:

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

Using the linearity of Expectation we get:

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

Notice that  $\rho(aX+b, cY+d) = \rho(X, Y)$ . This new quantity is independent of the change in scale and its value is quite informative.

# Covariance and Correlation

Properties of correlation:

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

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

$$\text{Corr}(X, Y) = \text{Cov}(X^*, Y^*) = E(X^*Y^*)$$

# Covariance and Correlation

**Claim:** The correlation is always between **-1** and **+1**

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

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

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

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

$$-1 \leq \text{Corr}(X, Y) \leq 1$$

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

# Correlation and Independence

$X$  &  $Y$  are uncorrelated iff any of the following hold

$$\text{Cov}(X, Y) = 0,$$

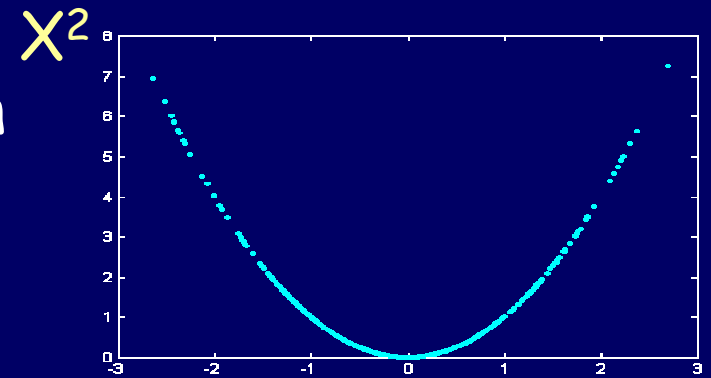
$$\text{Corr}(X, Y) = 0$$

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

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

**Example:** Let  $X \sim N(0,1)$  and  $Y = X^2$ , then

$\text{Cov}(XY) = E(XY) - E(X)E(Y) = E(X^3) = 0$ ,  
since the density is symmetric.



X

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:

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

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

$$\text{Var}(X+Y) = (p_1+p_2)(1 - p_1+p_2) N.$$

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

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

# Correlations in the Multinomial Distribution

Hence

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

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

$$\begin{aligned}\rho &= \frac{-N p_1 p_2}{\sqrt{N p_1 (1 - p_1)} \sqrt{N p_2 (1 - p_2)}} \\ &= \sqrt{\frac{p_1 p_2}{(1 - p_1)(1 - p_2)}}\end{aligned}$$

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

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

Proof:

$$\text{Var}(\sum_i X_i) = E[\sum_i X_i - E(\sum_j X_j)]^2$$

$$\begin{aligned} [\sum_i X_i - E(\sum_j X_j)]^2 &= [\sum_i (X_i - \mu_i)]^2 \\ &= \sum_i (X_i - \mu_i)^2 + 2 \sum_{j < i} (X_i - \mu_i)(X_j - \mu_j). \end{aligned}$$

Now take expectations and we have the result.

## Variance of the Sample Average

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

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x(i) \quad \& \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x(i) - \bar{x})^2$$

are the population mean and population variance.

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

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

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

## Variance of the Sample Average

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

When  $X_1, X_2, \dots, X_n$  are drawn with replacement, they are independent and each  $X_k$  has variance  $\sigma^2$ . Then

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

## Variance of the Sample Average

**Question:** What is the SD for sampling without replacement?

**Solution:** Let  $S_n = X_1 + X_2 + \dots + X_n$ . Then  $\bar{X}_n = S_n / n$ .

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

By symmetry  $\text{Cov}(X_i, X_j) = \text{Cov}(X_1, X_2)$ , so

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

This formula hold for all  $2 \leq n \leq N$ .

When  $n=N$   $\text{Var}(S_N)=0$  and  $S_N/N = \bar{x}$  -- the sample is the entire population drawn out in random order. However,  $\text{Cov}(X_1, X_2)$  should not depend on the ultimate sample size, so we use the formula with  $n=N$  and obtain:

$$\text{Cov}(X_1 X_2) = -\sigma^2 / (N-1).$$

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

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