

Slides for Data Relationships

MSIT 431
Week #8

Example of Two Related Random Variables

- X = Education level in years
 - Maybe subdivided into separate ranges
 - We shall use 6 ranges in our example
- Y = Household income per year in \$1,000
 - Maybe subdivided into separate ranges
 - We shall use 7 ranges in our example

Discrete Ranges for X

- $A1 = \{X=6, \text{No than High School}\}$
- $A2 = \{X = 10, \text{some High School}\}$
- $A3 = \{X =12, \text{High School}\}$
- $A4 = \{X=14, \text{some Colleg}\}$
- $A5=\{X=16, \text{BS Degree}\}$
- $A6=\{X=20, \text{Higher Degree}\}$
- We may have more levels if needed

Discrete Ranges for Y

- $B1 = \{Y \text{ is less than } \$15k\}$
- $B2 = \{Y \text{ is between } \$15 \text{ to } \$30k\}$
- $B3 = \{Y \text{ is between } \$30 \text{ and } \$45k\}$
- $B4 = \{Y \text{ is between } \$45 \text{ and } \$75k\}$
- $B5 = \{Y \text{ is between } \$75 \text{ and } \$150k\}$
- $B6 = \{Y \text{ is between } \$150 \text{ and } \$300k\}$
- $B7 = \{Y \text{ is } \$300k \text{ or higher}\}$

Raw Data Relating X and Y

No.	A1	A2	A3	A4	A5	A6	Total
B1	957	1468	4179	2831	1165	324	10924
B2	889	1425	6762	5057	2219	583	16935
B3	215	528	3953	4059	2797	1133	12685
B4	79	198	2203	2752	2961	1786	9979
B5	11	40	463	858	1478	1275	4125
B6	4	6	79	121	327	461	998
B7	1	3	7	10	10	26	57
Total	2156	3668	17646	15688	10957	5588	55703

Joint and Marginal Probabilities

P(i,j)	A1	A2	A3	A4	A5	A6	P(Bj)
B1	0.01718	0.026354	0.075023	0.050823	0.020914	0.005817	0.196112
B2	0.01596	0.025582	0.121394	0.090785	0.039836	0.010466	0.304023
B3	0.00386	0.009479	0.070966	0.072869	0.050213	0.02034	0.227726
B4	0.001418	0.003555	0.039549	0.049405	0.053157	0.032063	0.179147
B5	0.000197	0.000718	0.008312	0.015403	0.026534	0.022889	0.074053
B6	7.18E-05	0.000108	0.001418	0.002172	0.00587	0.008276	0.017916
B7	1.8E-05	5.39E-05	0.000126	0.00018	0.00018	0.000467	0.001023
P(Ai)	0.038705	0.065849	0.316787	0.281637	0.196704	0.100318	1

Conditional Probabilities of Y Given X

$P(B_j A_i)$	A1	A2	A3	A4	A5	A6	$P(B_j)$
B1	0.443878	0.400218	0.236824	0.180456	0.106325	0.057981	0.196112
B2	0.412338	0.388495	0.383203	0.322348	0.202519	0.104331	0.304023
B3	0.099722	0.143948	0.224017	0.258733	0.255271	0.202756	0.227726
B4	0.036642	0.05398	0.124844	0.175421	0.270238	0.319613	0.179147
B5	0.005102	0.010905	0.026238	0.054691	0.134891	0.228168	0.074053
B6	0.001855	0.001636	0.004477	0.007713	0.029844	0.082498	0.017916
B7	0.000464	0.000818	0.000397	0.000637	0.000913	0.004653	0.001023
Total	1	1	1	1	1	1	1
$P(A_i)$	0.038705	0.065849	0.316787	0.281637	0.196704	0.100318	1

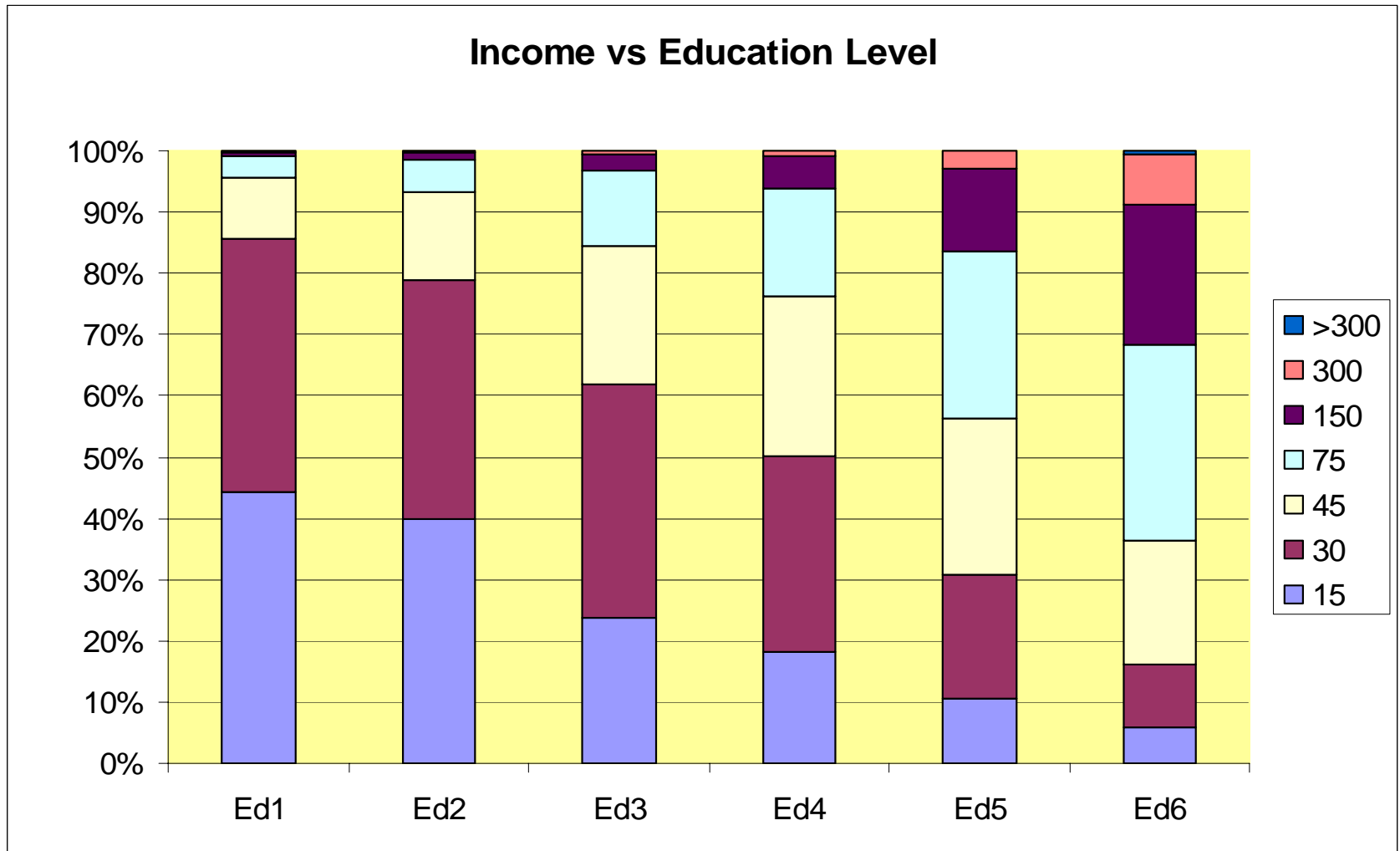
Conditional Probabilities of X Given Y

$P(A_i B_j)$	A1	A2	A3	A4	A5	A6	Total
B1	0.087605	0.134383	0.382552	0.259154	0.106646	0.029659	1
B2	0.052495	0.084145	0.399291	0.298612	0.13103	0.034426	1
B3	0.016949	0.041624	0.311628	0.319984	0.220497	0.089318	1
B4	0.007917	0.019842	0.220764	0.275779	0.296723	0.178976	1
B5	0.002667	0.009697	0.112242	0.208	0.358303	0.309091	1
B6	0.004008	0.006012	0.079158	0.121242	0.327655	0.461924	1
B7	0.017544	0.052632	0.122807	0.175439	0.175439	0.45614	1

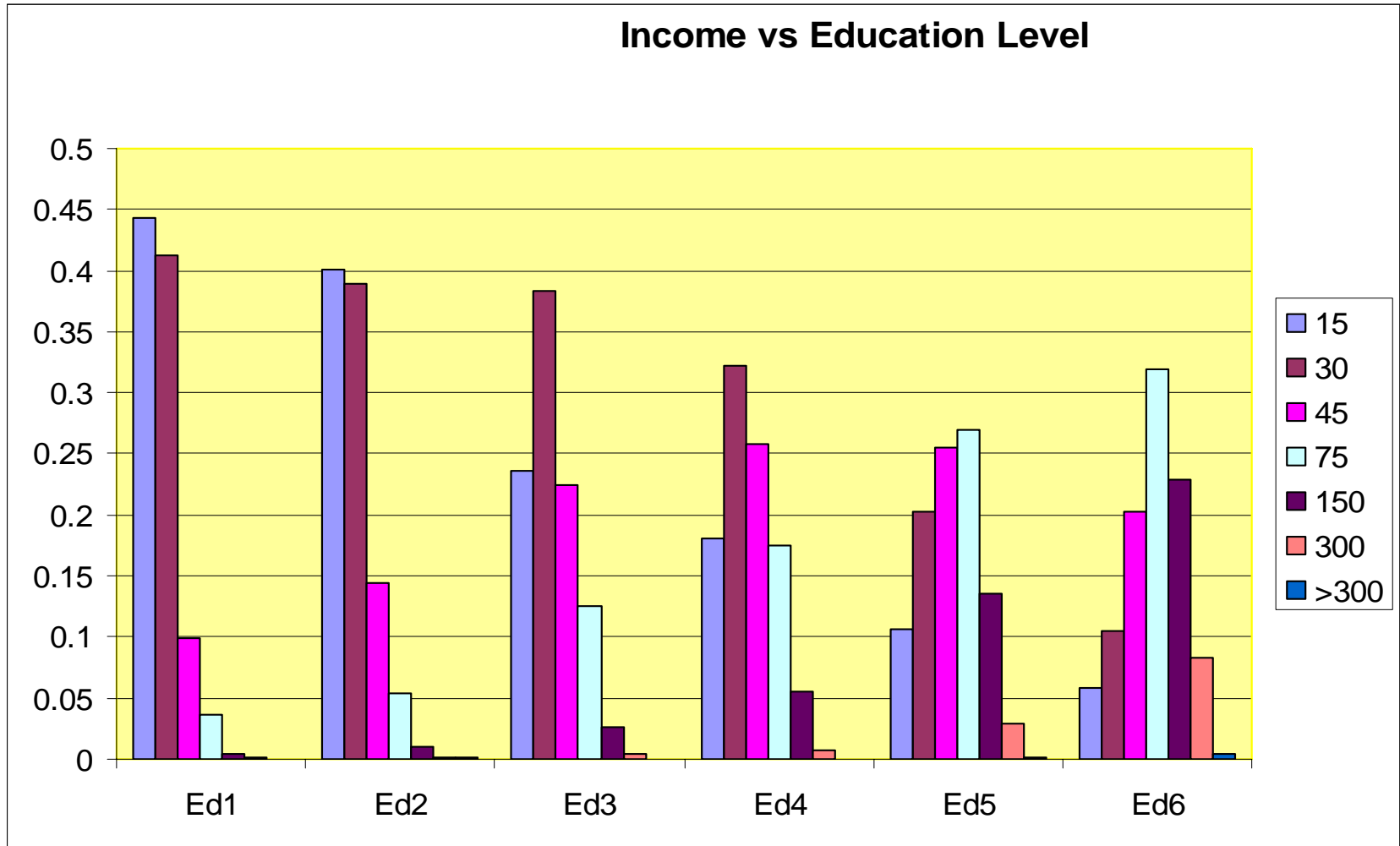
Joint and Marginal PMF

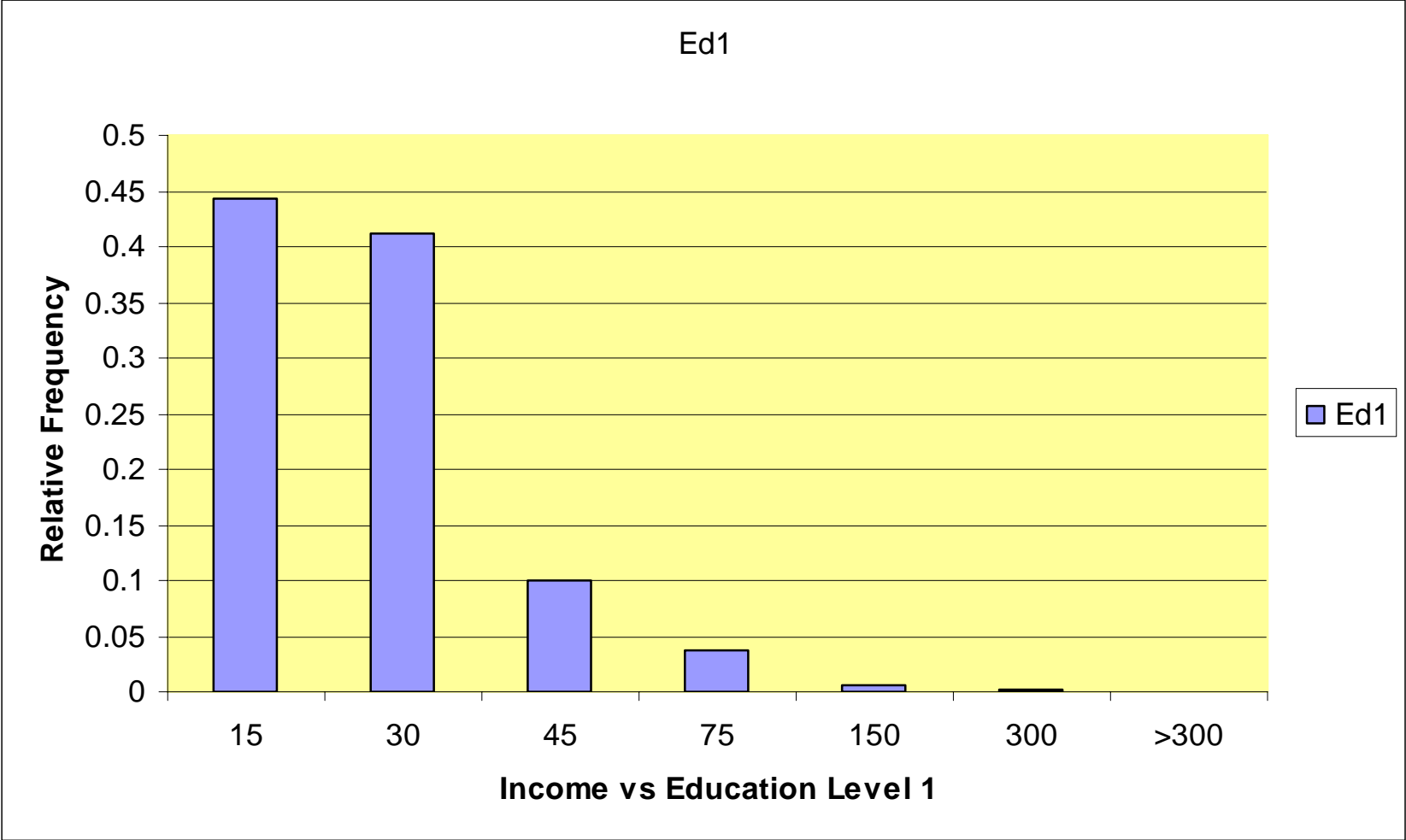
- $P(A_i, B_j) = \text{Probability (X in } A_i \text{ and Y in } B_j)$
- $P(A_i) = \text{Probability (X in } A_i) = \sum_{j=1}^7 P(A_i, B_j)$
- $P(B_j) = \text{Probability (Y in } B_j) = \sum_{i=1}^6 P(A_i, B_j)$
- $P(A_i|B_j) = P(A_i, B_j)/P(B_j) = P(B_j|A_i) P(A_i)/P(B_j)$
- $P(B_j) = \sum_{i=1}^6 P(B_j | A_i)P(A_i)$

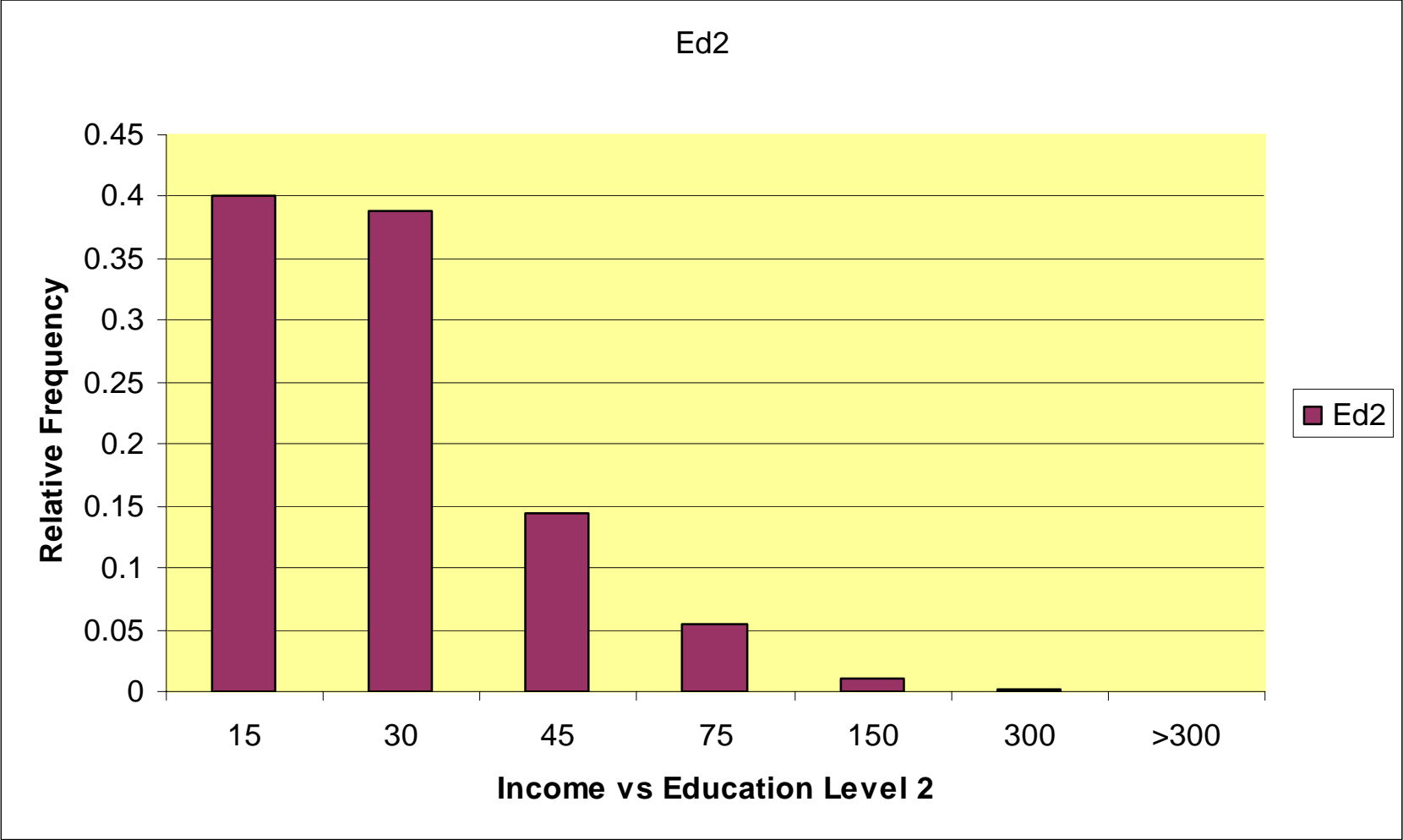
Conditional Probability of Y Given X

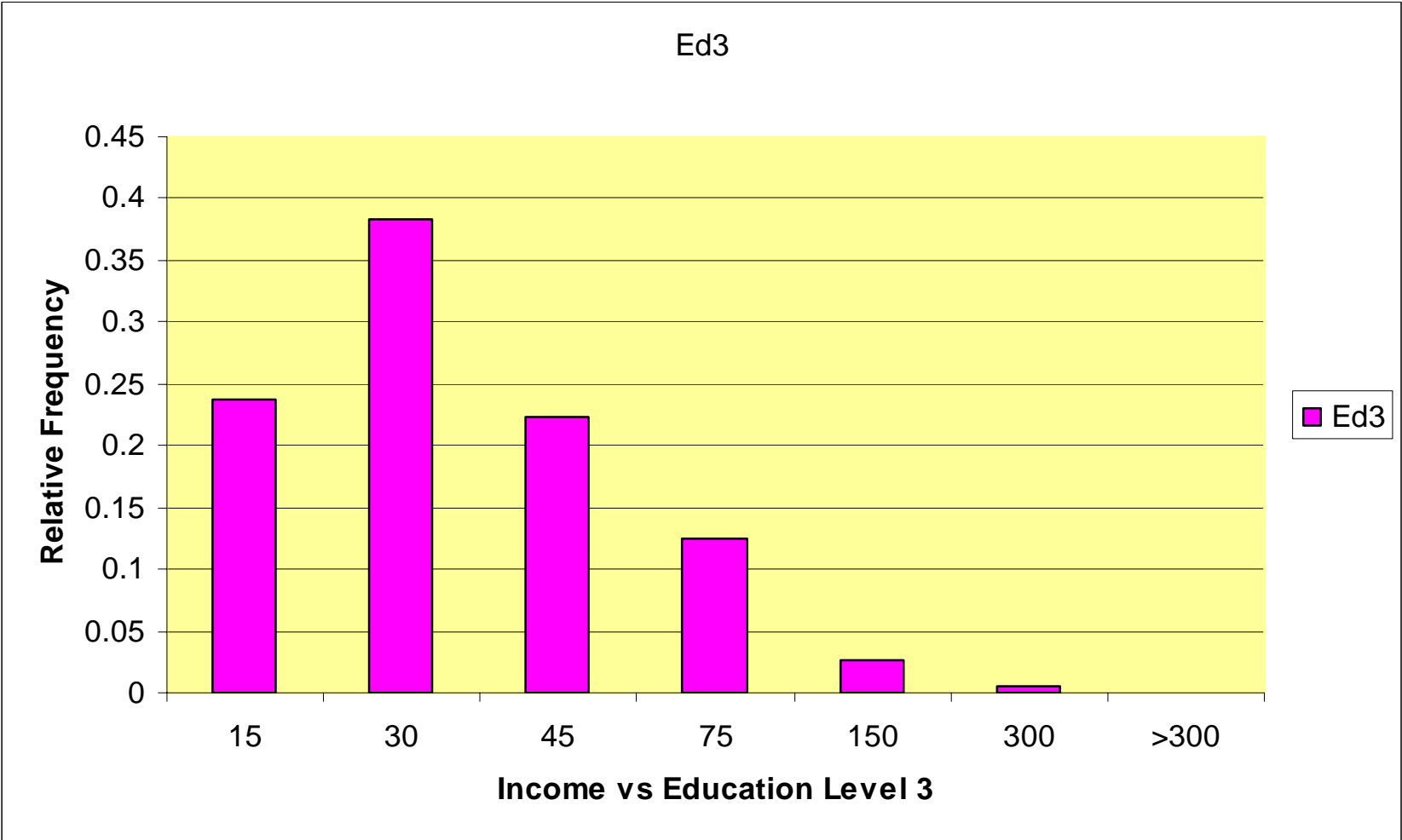


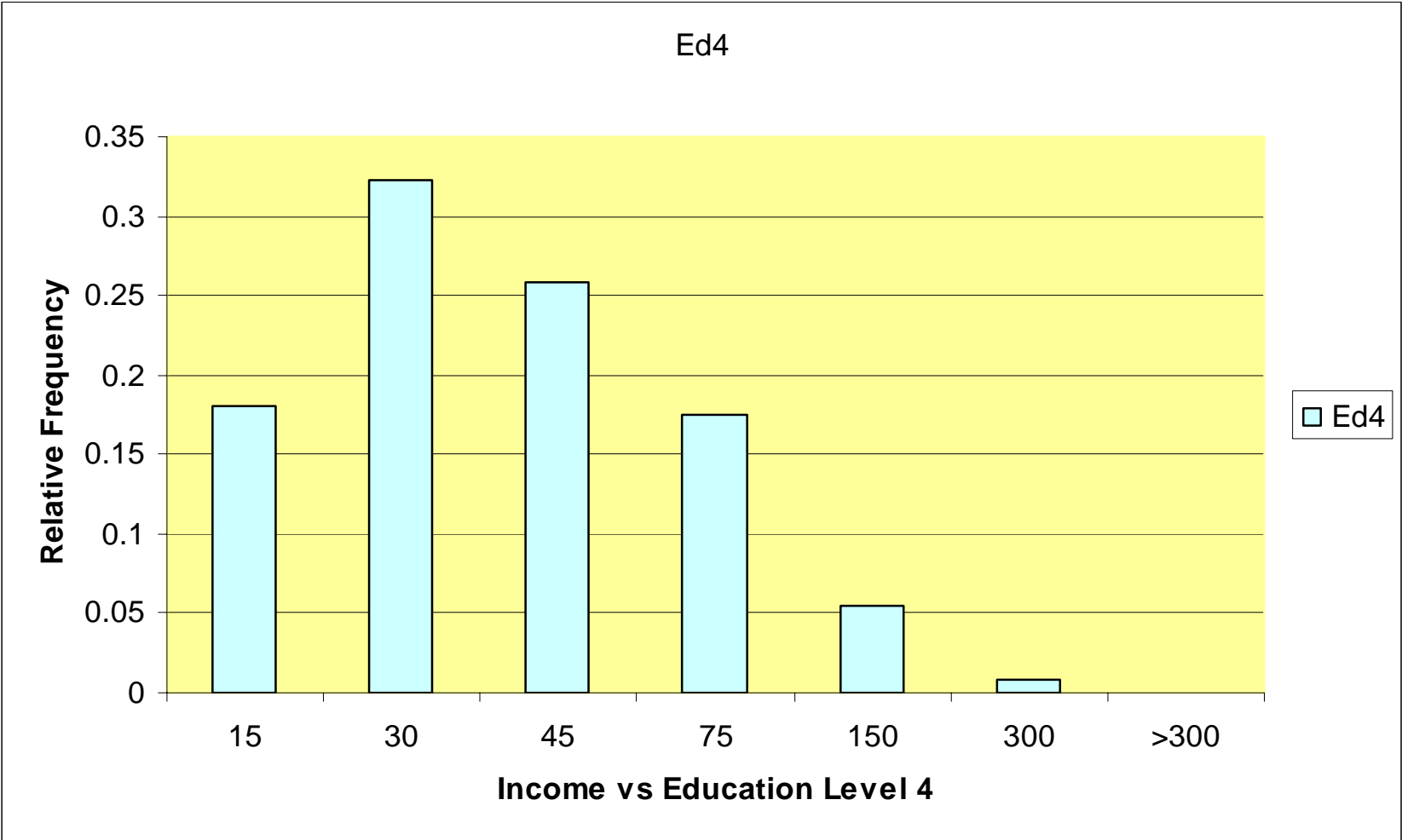
Conditiontional Probabilities of Y Given X

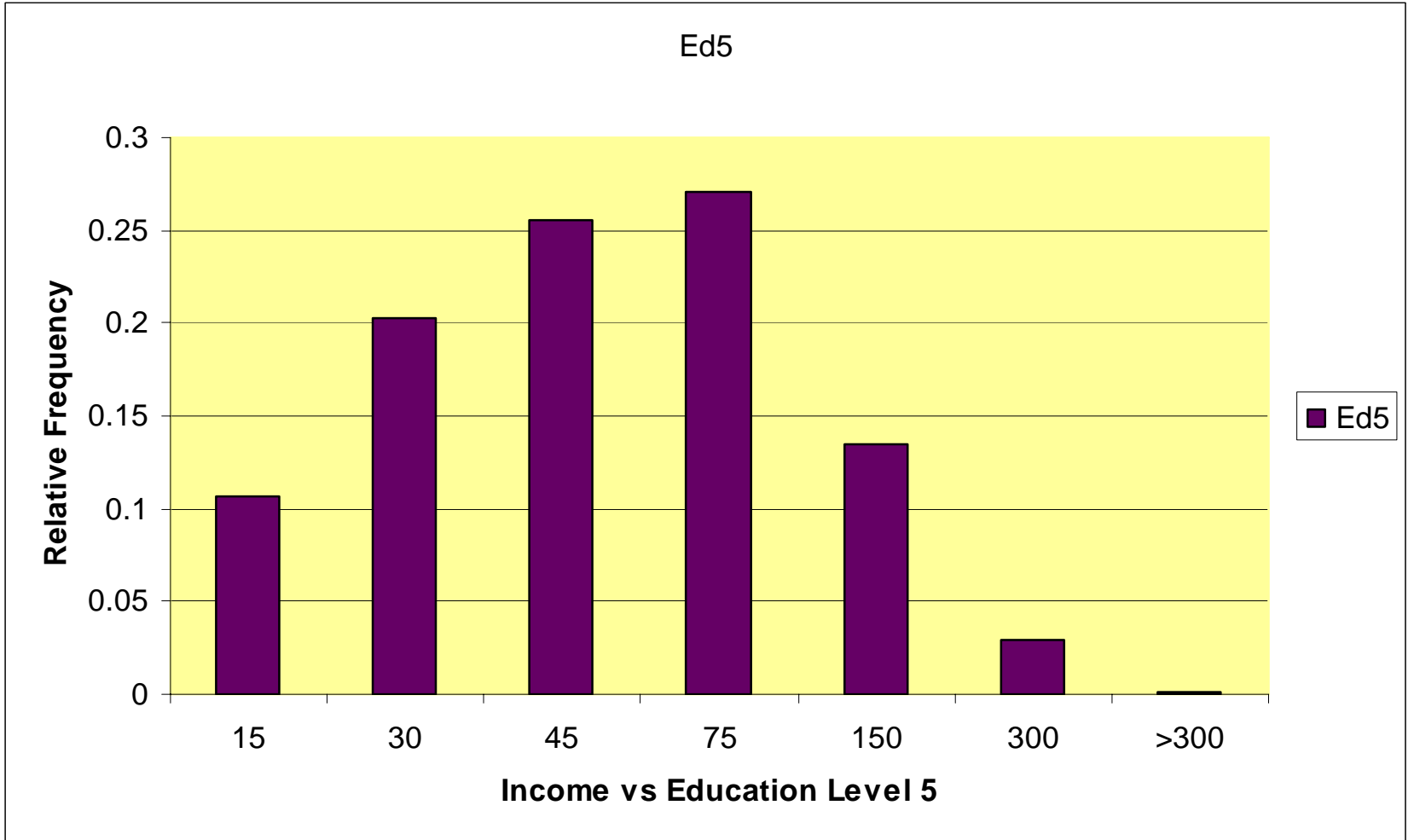


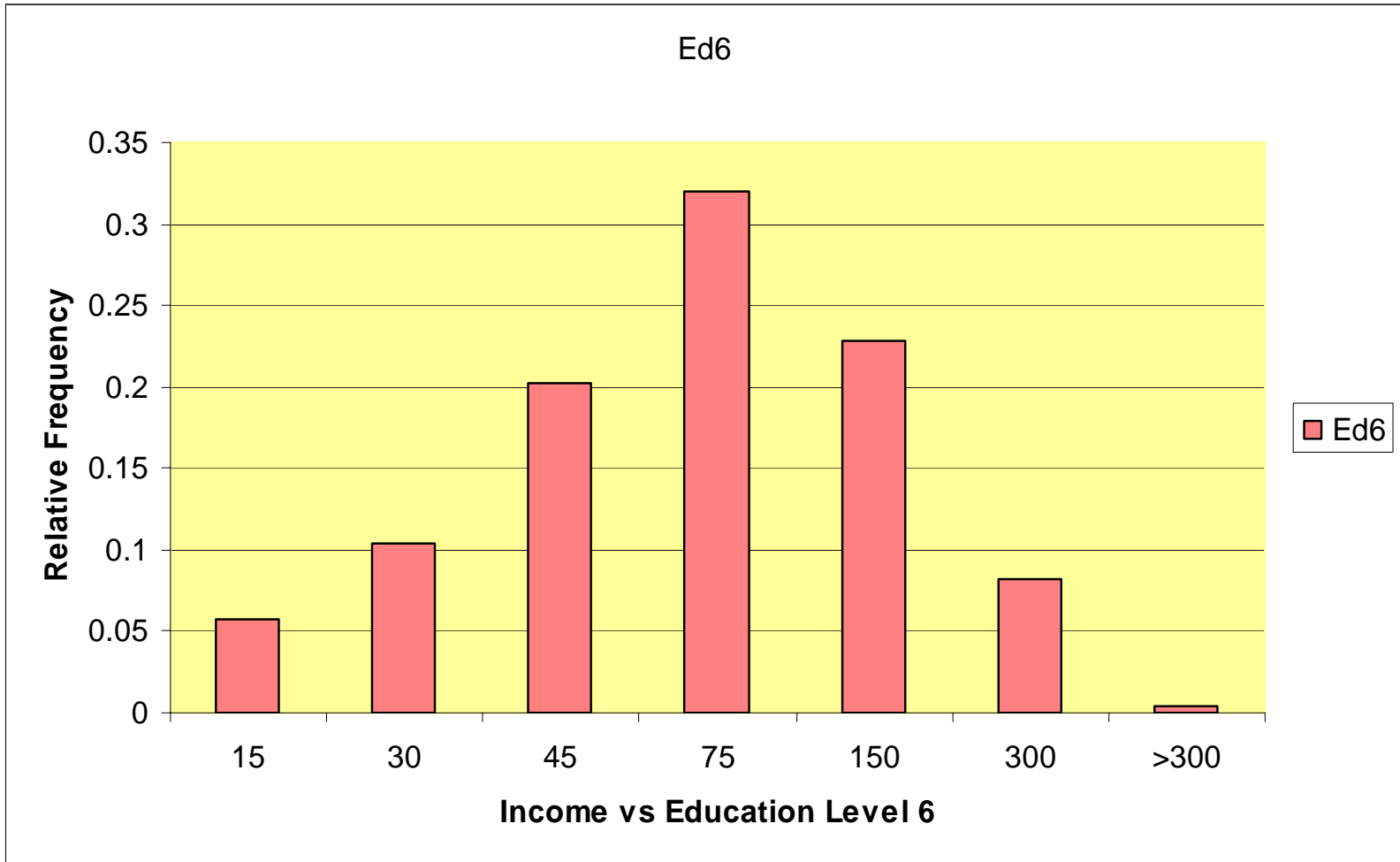




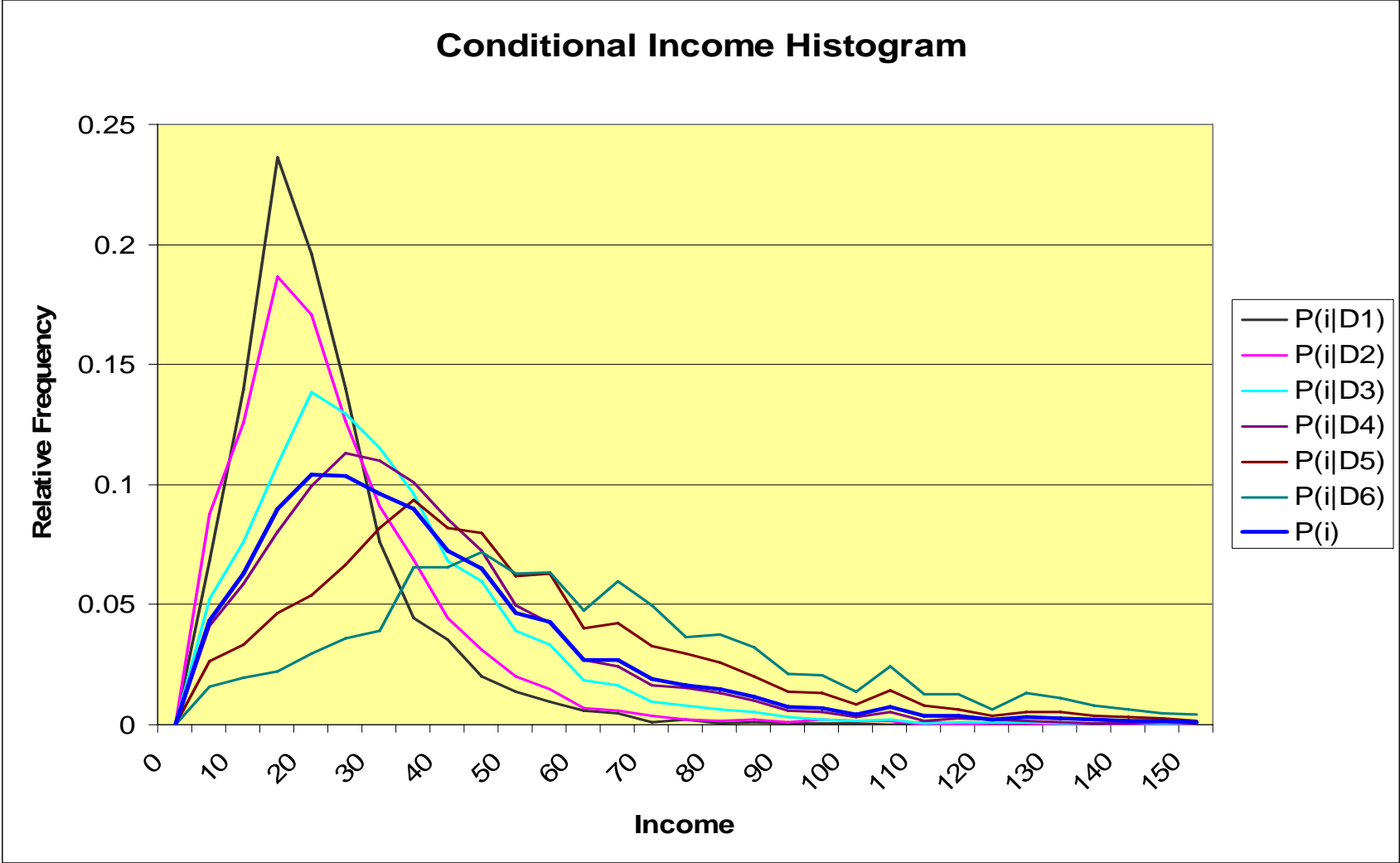




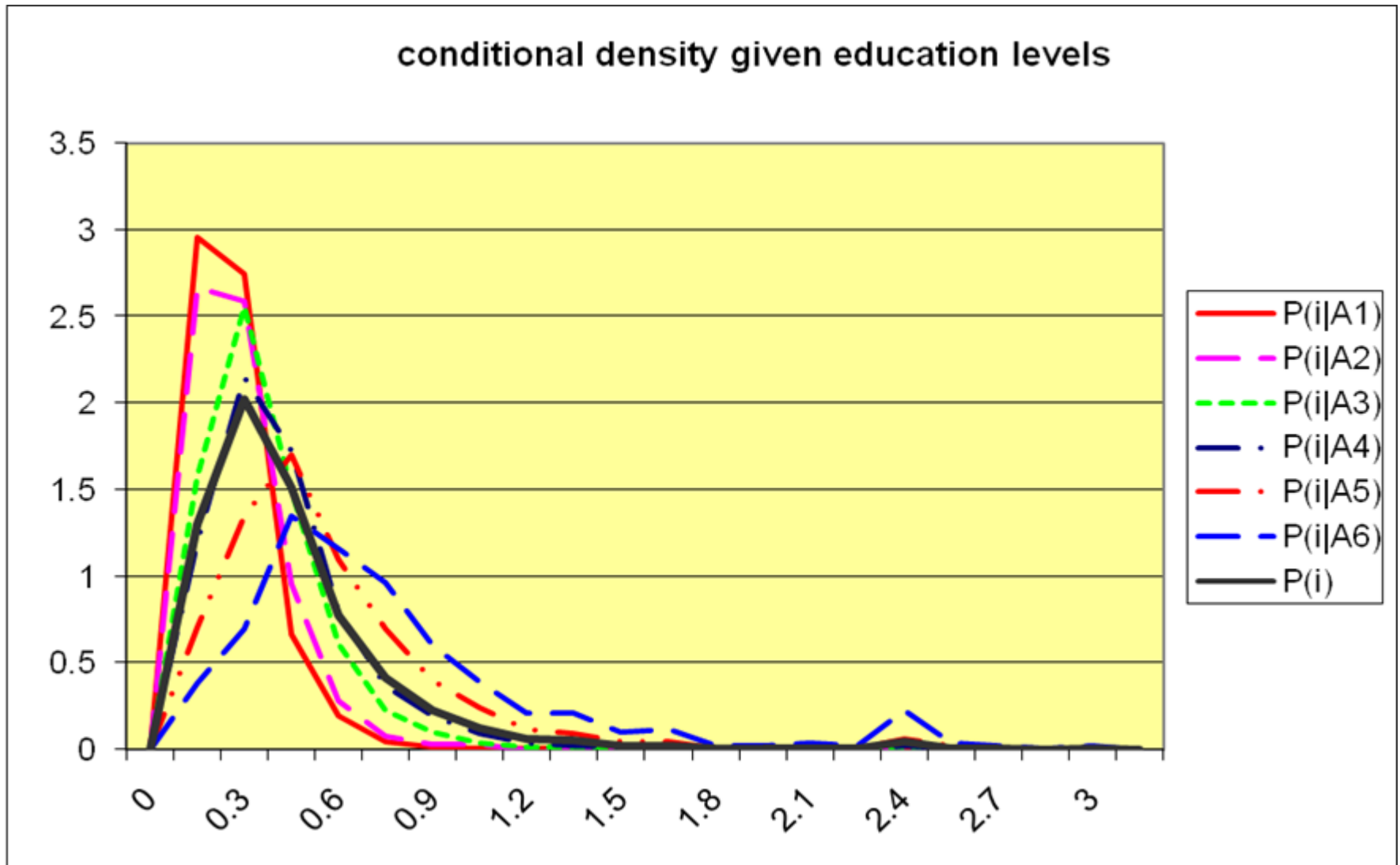




Conditional histograms of Y given education levels



Conditional densities of income given education levels



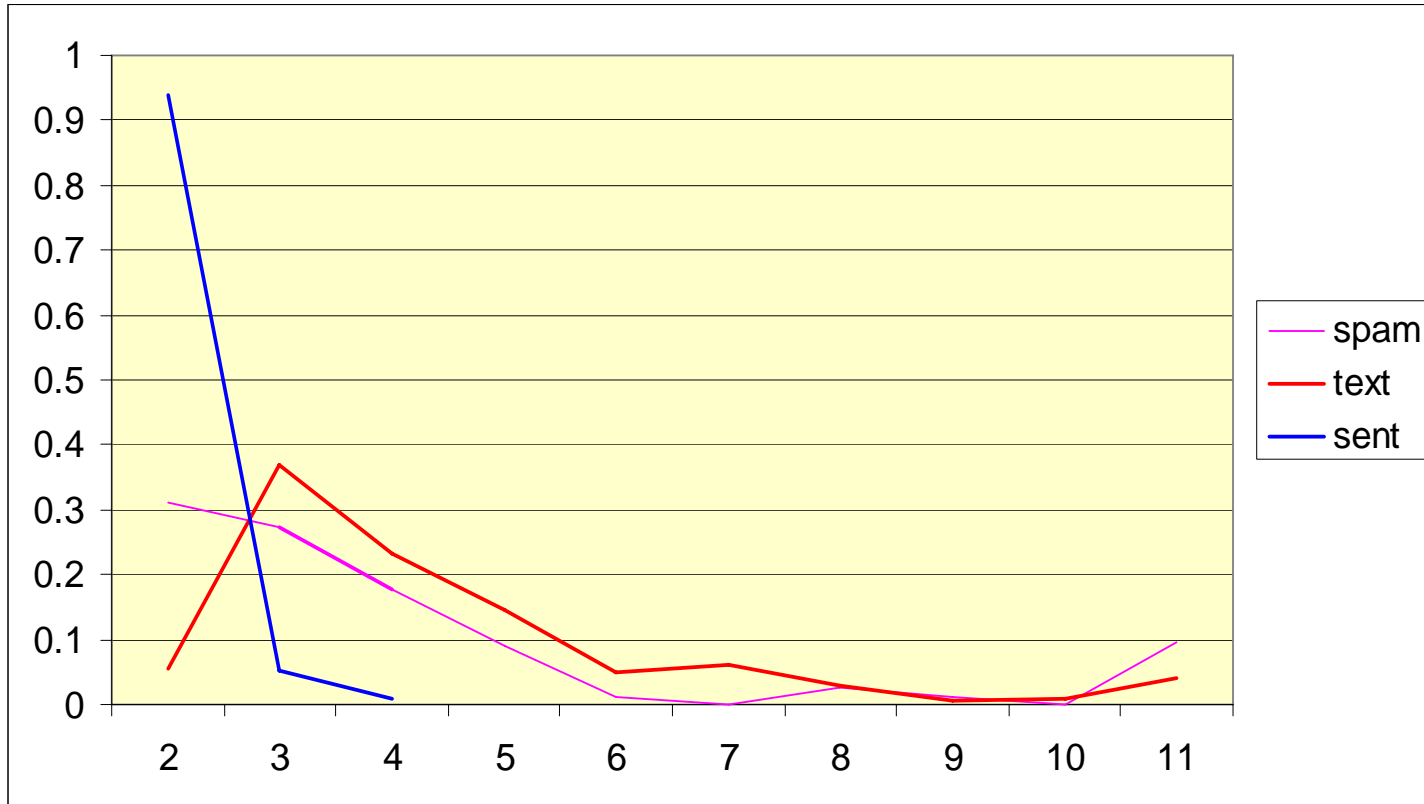
Conditional Probability Density Functions

- $f_{Y|X}(y|A_i)$ = Probability Density of Y given that X is in A_i
- $f_Y(y) = \sum_{i=1}^6 f_{Y|X}(y|A_i)P(A_i)$
= Unconditional Density
- $P(A_i|Y=y) = f(y|A_i)P(A_i)/f_Y(y)$ = Bayes' Rule

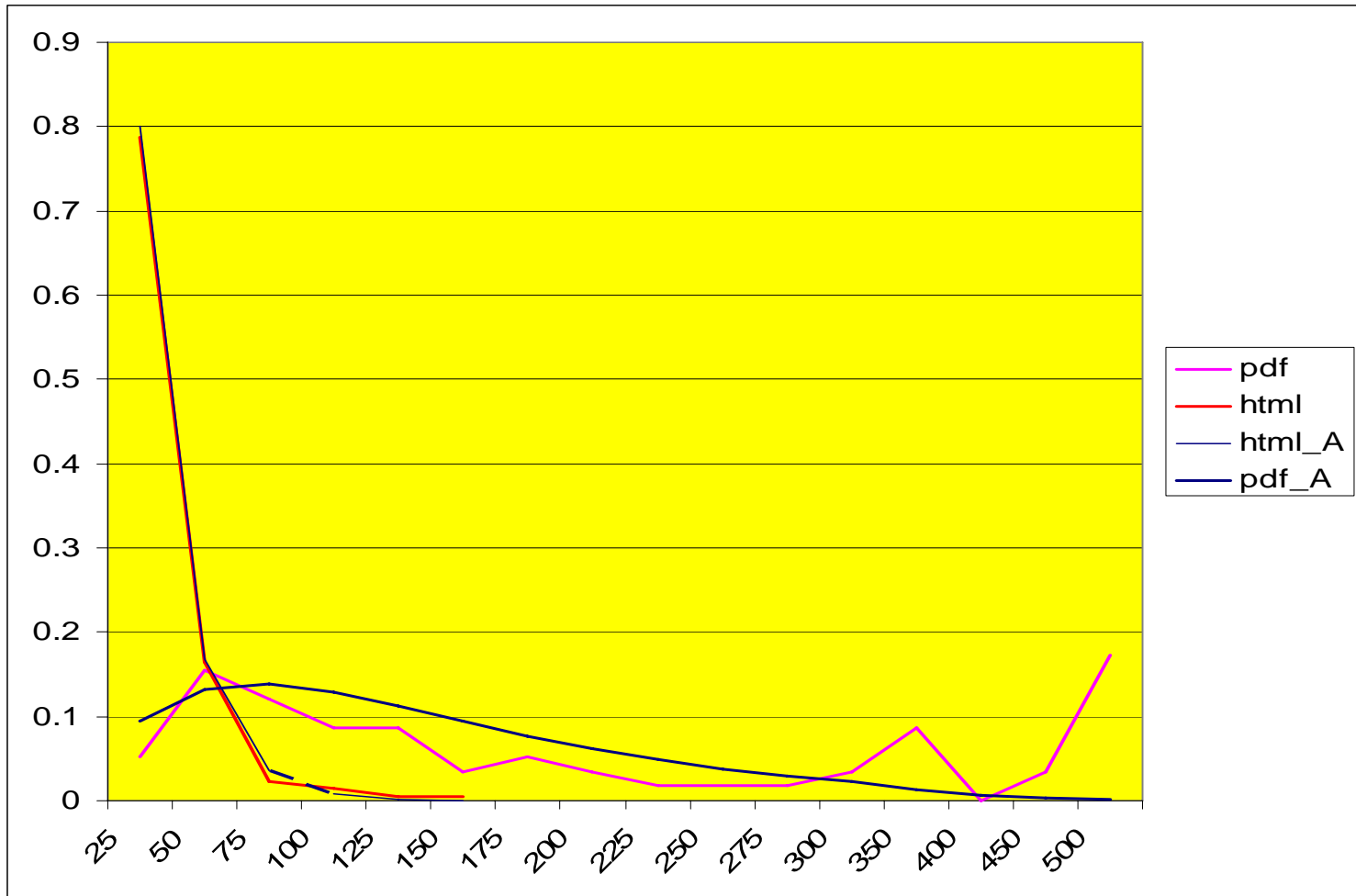
Message Length Y by Type X

- $X=1$ = messages that have html attachment
- $X=2$ = messages that have pdf or doc attachments
- $X=3$ = messages that contain text only
- $X=4$ = messages that eluded my spam catcher
- $X=5$ = copies of messages mailed from my home machine (text only, short header)

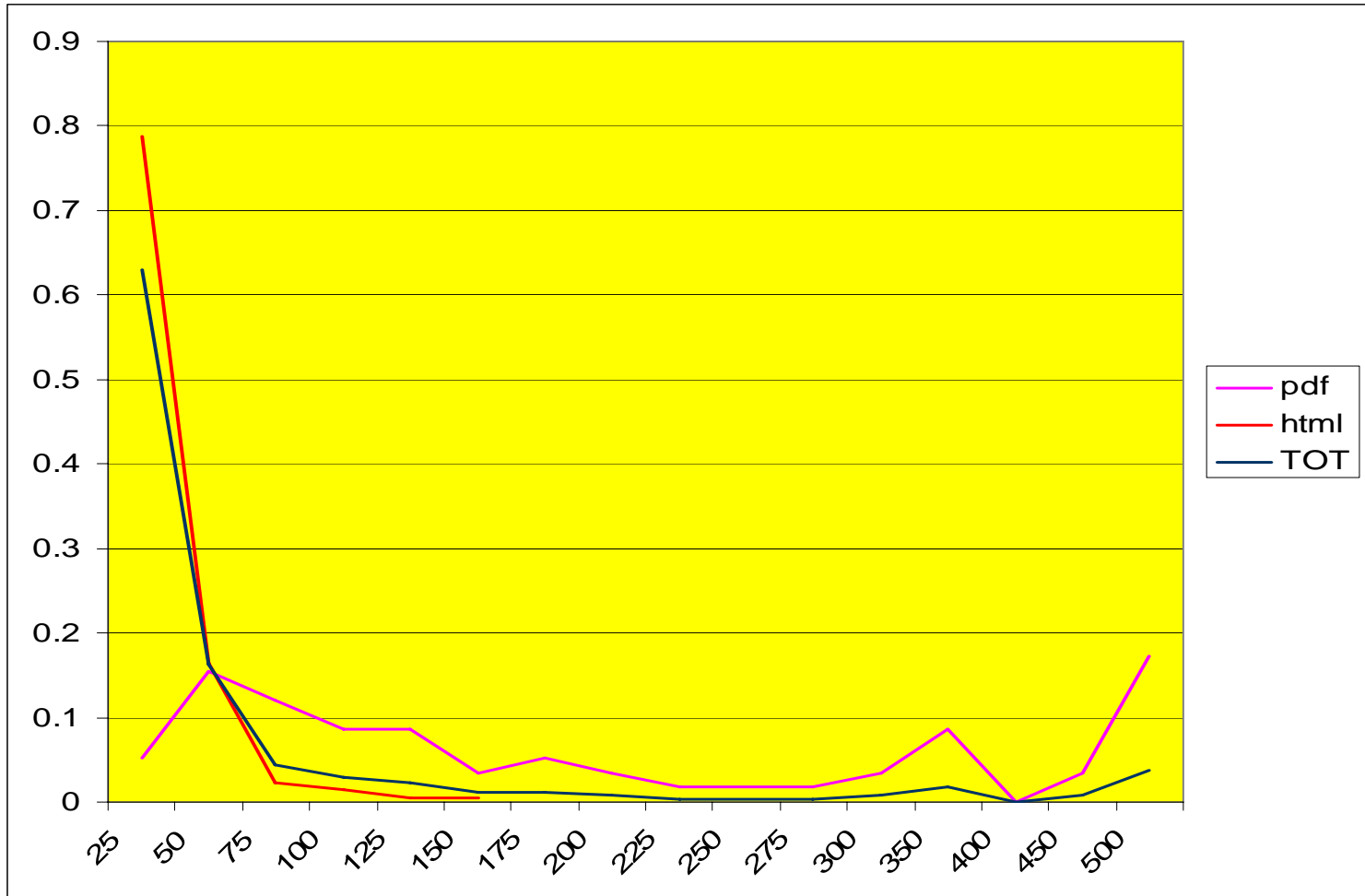
Conditional densities of message size for $X=3$ (text), $X=4$ (spam), and $X=5$ (sent)



Conditional densities of message size for $X=1$ (html) and $X=2$ (pdf), together with their analytical approximations (A)



Conditional and unconditional densities of message lengths involving only $X=1$ (html) and $X=2$ (pdf)



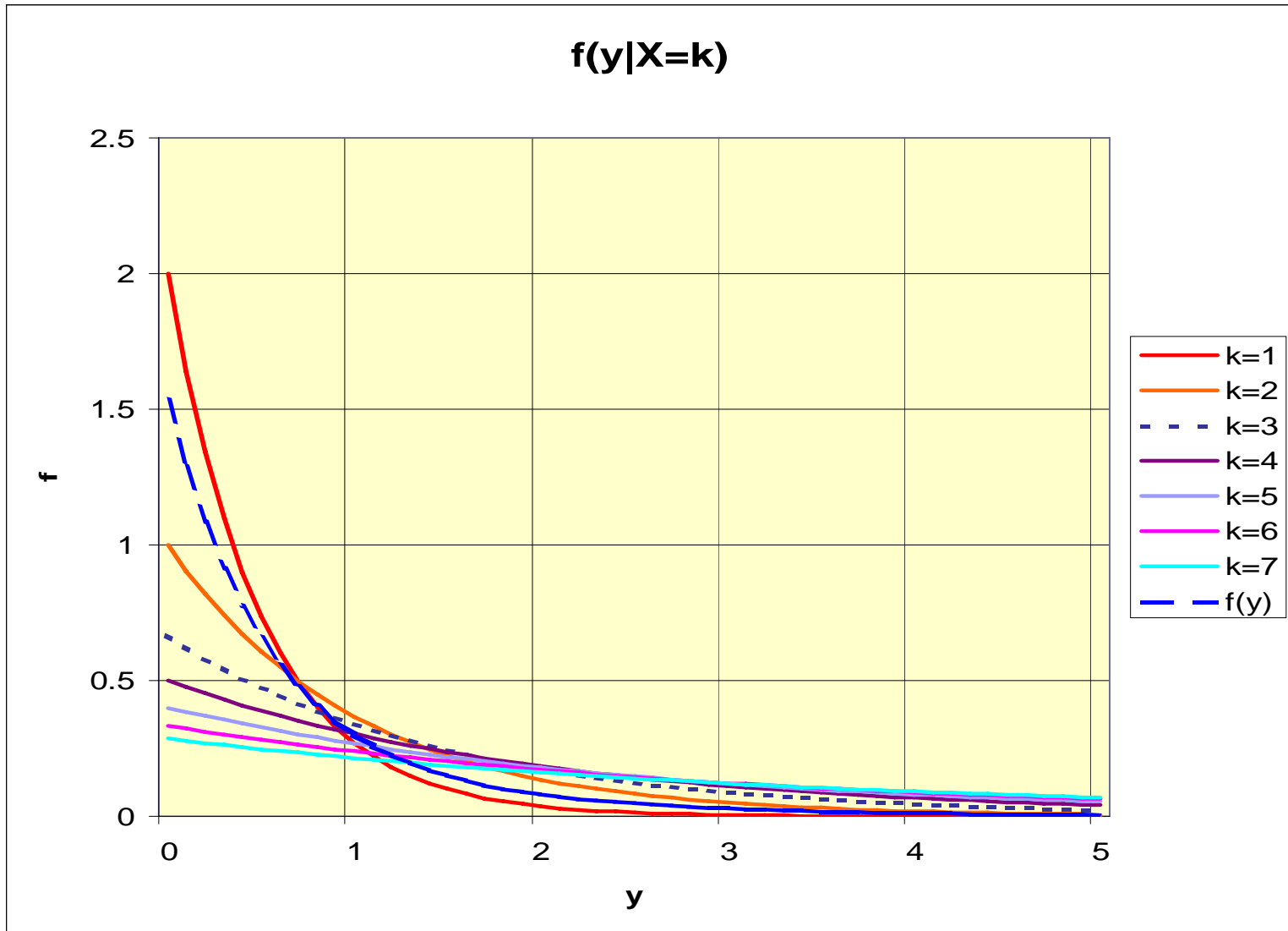
Ex: 5.1 Delay Y vs # of Tx

- $P\{X = k\} = p (1 - p)^{k-1}$, for $k = 1, 2, 3, \dots$
- $f_{Y|X}(y | X = k) = \{\mu/k\} \exp\{-\mu y/k\}$
- $E\{Y|X=k\} = k/\mu = \text{Conditional Mean}$
- $f_Y(y) = \sum_{k=1}^{\infty} f_{Y|X}(y | X = k) P\{X = k\} = \sum_{k=1}^{\infty} (\mu/k) \exp(-\mu y/k) p(1-p)^{k-1}$
- $E\{Y\} = 1/(p\mu)$
- $P(X = k | Y = y) = f_{Y|X}(y | X = k) P\{X=k\}/f_Y(y)$

Conditional Densities & Probabilities

- Compute for $p=0.6$ and $\mu=2$
- $P\{X=k\}=0.6(0.4)^{k-1}$; $f_{Y|X}(y|X=k)=(2/k)\exp(-2y/k)$
- $P\{a<Y<b|X=k\} = \exp(-\mu a/k) - \exp(-\mu b/k)$
 $= \exp\{-2a/k\} - \exp\{-2b/k\}$
- $P\{a<Y<b\}=\sum\{\exp(-2a/k)-\exp(-2b/k)\}0.6(0.4)^{k-1}$
- $P\{X=k|a<Y<b\} = P\{a<Y<b|X=k\}P\{X=k\}/P\{a<Y<b\}$
- Numerical results shown for $a=0$, $b=1$

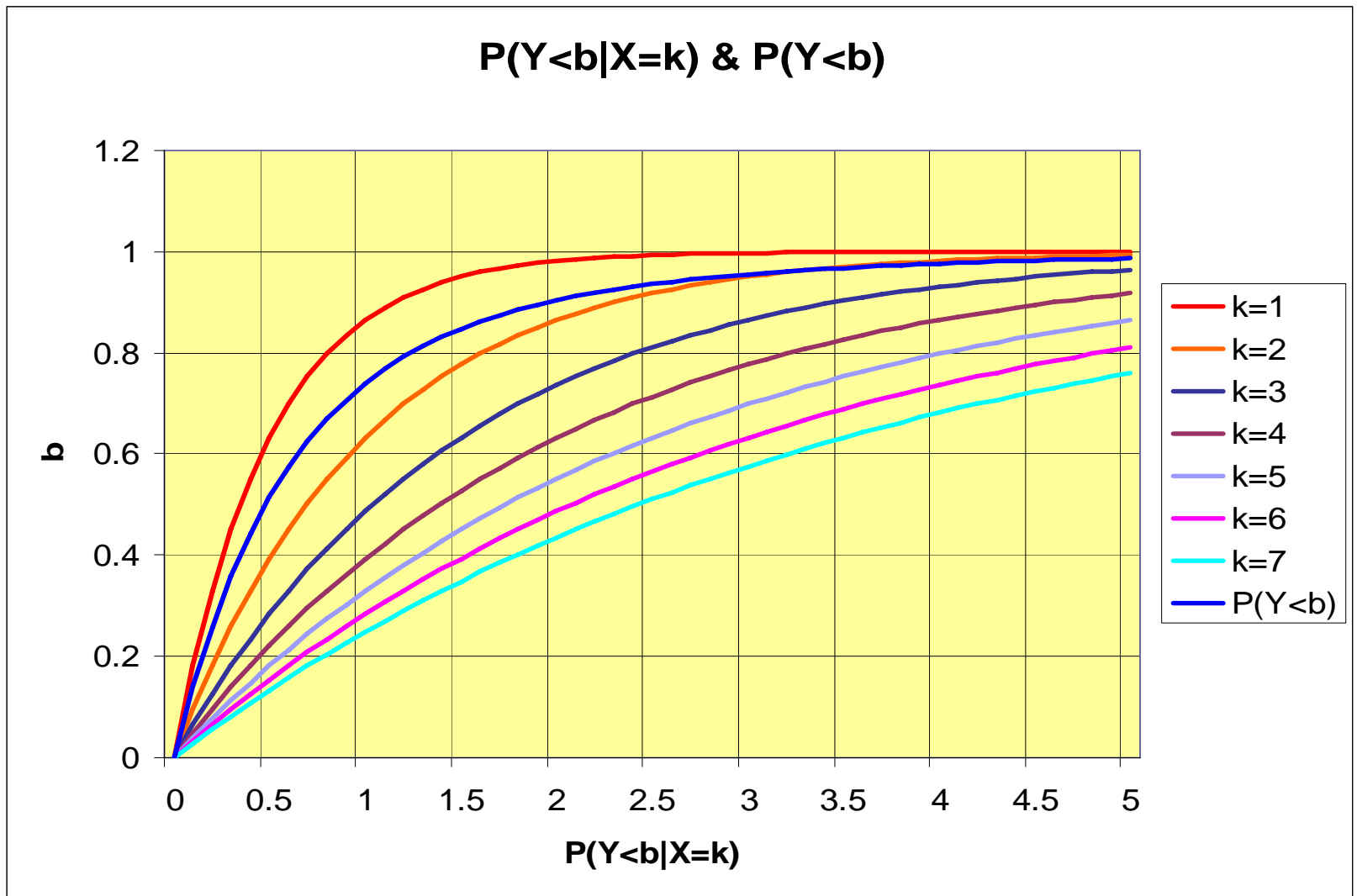
Conditional densities for the example



Conditional probabilities & Bayes' rule

	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9		
$P\{Y<1 X=k\}$	0.8647	0.6321	0.4866	0.3935	0.3297	0.2835	0.24852	0.2212	0.1993		
										Sum of $P\{X=k\}$	
$P\{X=k\}$	0.6	0.24	0.096	0.0384	0.0154	0.0061	0.00246	0.00098	0.0004	0.99974	
										$P\{Y<1\}$	
$P\{X=k, Y<1\}$	0.5188	0.1517	0.0467	0.0151	0.0051	0.0017	0.00061	0.00022	8E-05	0.74004	
										Sum of $P\{X=k Y<1\}$	
$P\{X=k Y<1\}$	0.701	0.205	0.0631	0.0204	0.0068	0.0024	0.00083	0.00029	0.0001	1	

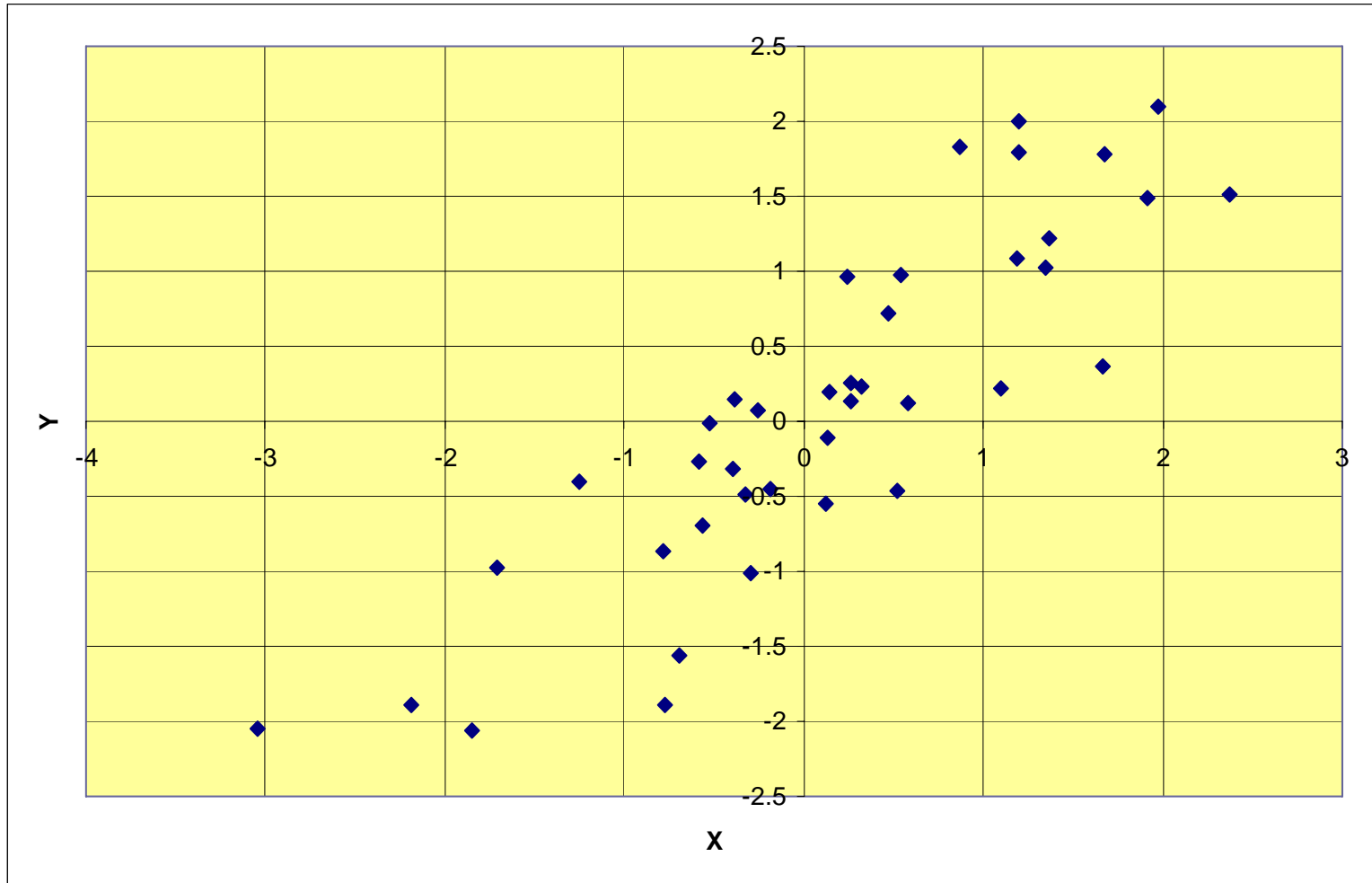
Conditional probabilities of $\{Y < b\}$



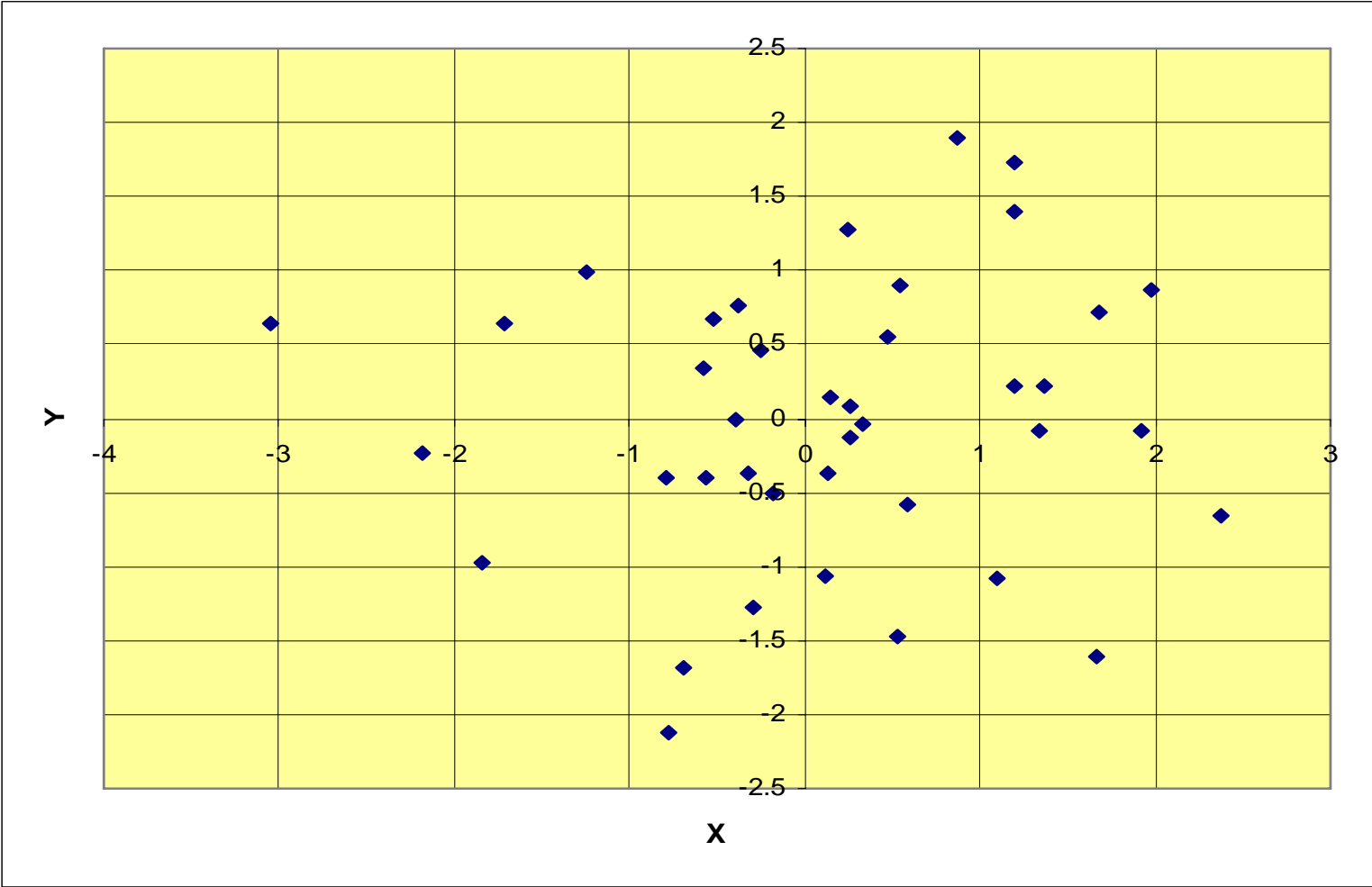
Summary on Presenting Relationships

- Stacked bar graphs
- A series of bar graphs
- Conditional box plots
- Conditional probabilities
- Conditional densities
- Joint probabilities
- Scatter diagrams

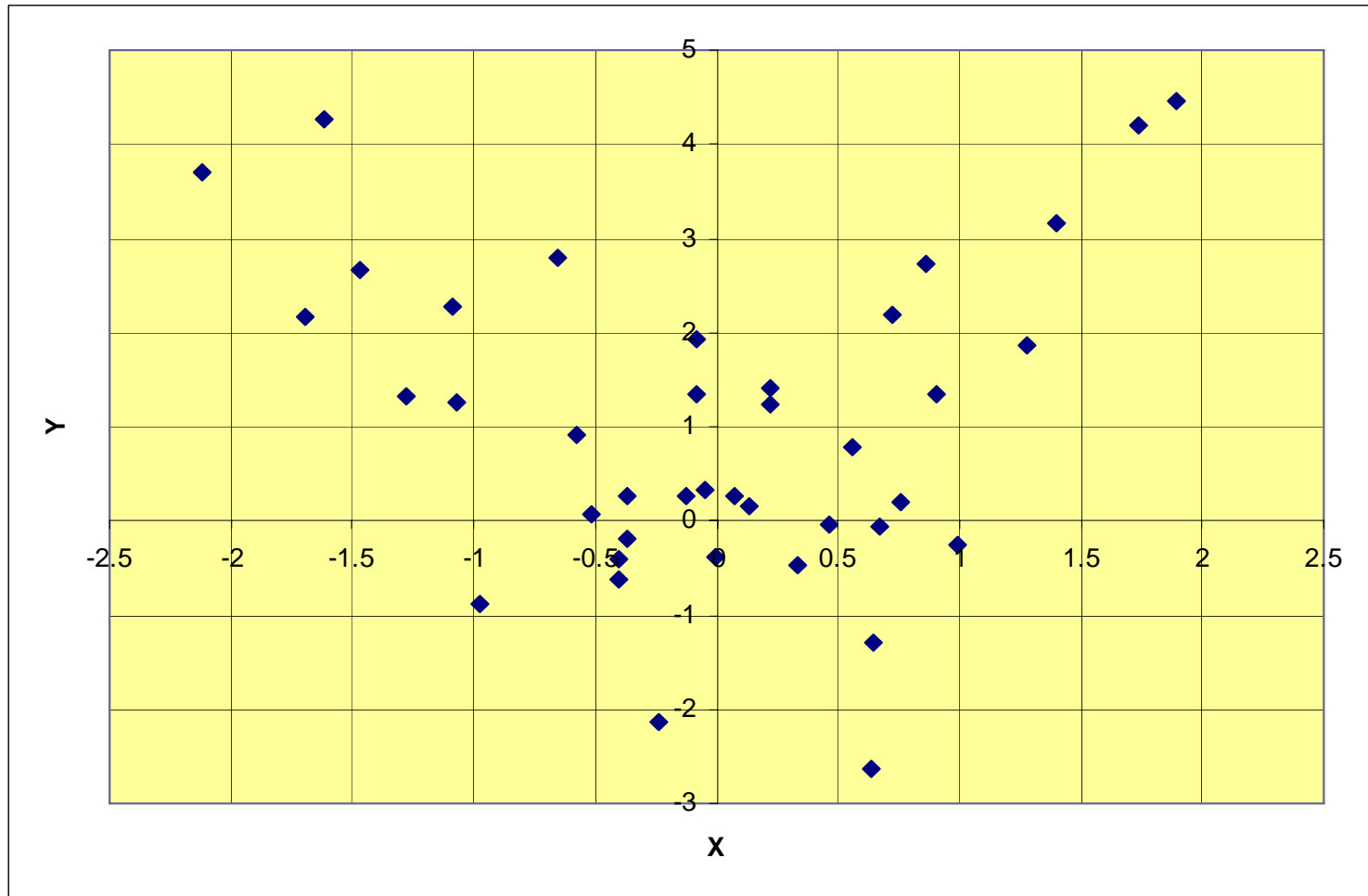
Scatter Diagrams 5.5(a)



Scatter Diagrams 5.5(b)



Scatter Diagrams 5.6



Correlation from Data

- Sample means: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$
- Sample variances: $s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ $s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$
- Sample correlation: $r = \frac{1}{(n-1)s_x s_y} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$
- The value of r: $-1 \leq r \leq 1$

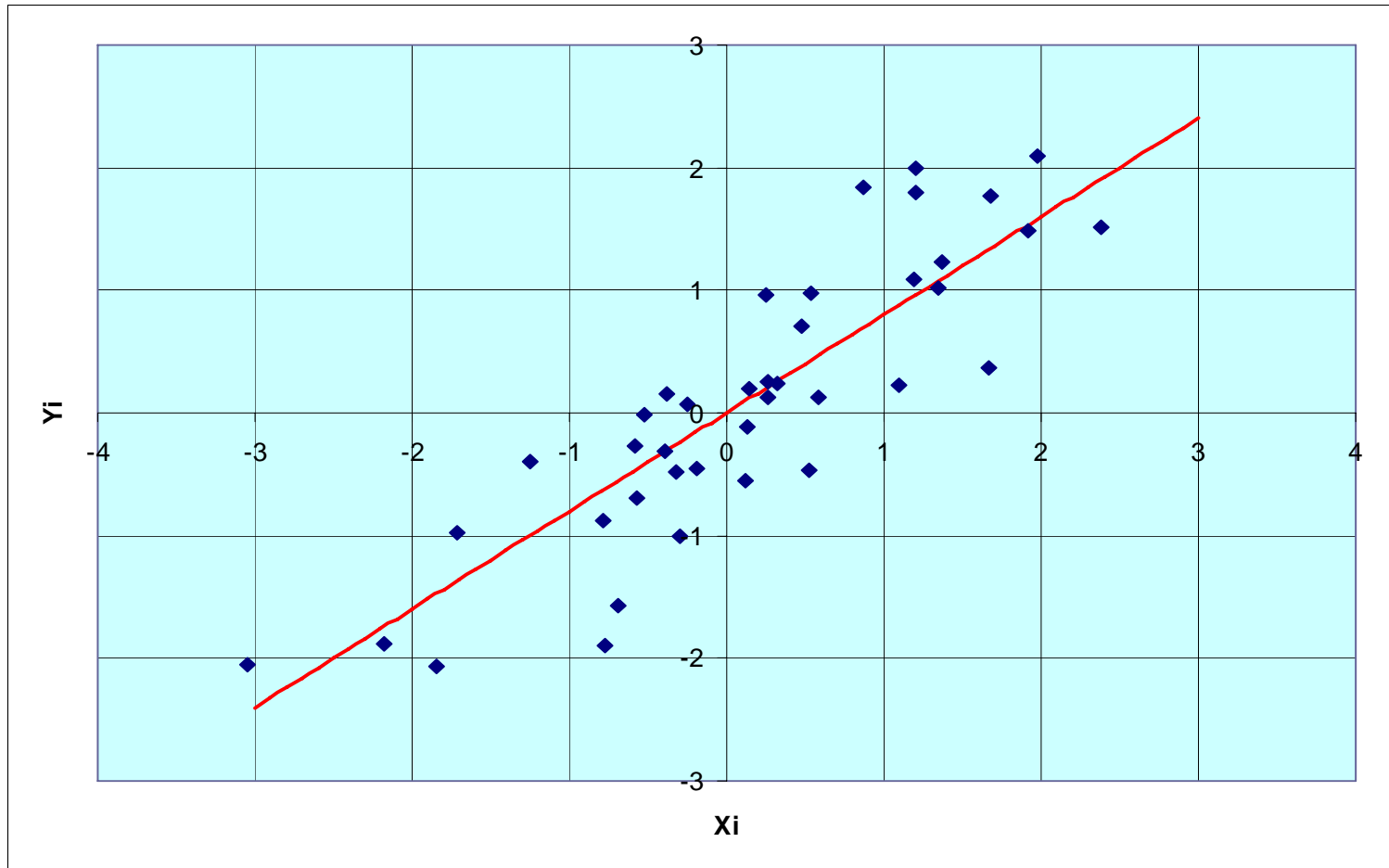
Properties of Correlation

- $r = 0$: $\{x_i\}$ and $\{y_i\}$ are uncorrelated
- $r = \pm 1$: $\{x_i\}$ and $\{y_i\}$ are perfectly correlated
- $r > 0$: $\{x_i\}$ and $\{y_i\}$ are positively correlated
- $r < 0$: $\{x_i\}$ and $\{y_i\}$ are negatively correlated
- If r is close to 1 or -1 : $\{x_i\}$ and $\{y_i\}$ are strongly correlated
- If r is close to 0: $\{x_i\}$ and $\{y_i\}$ are weakly correlated

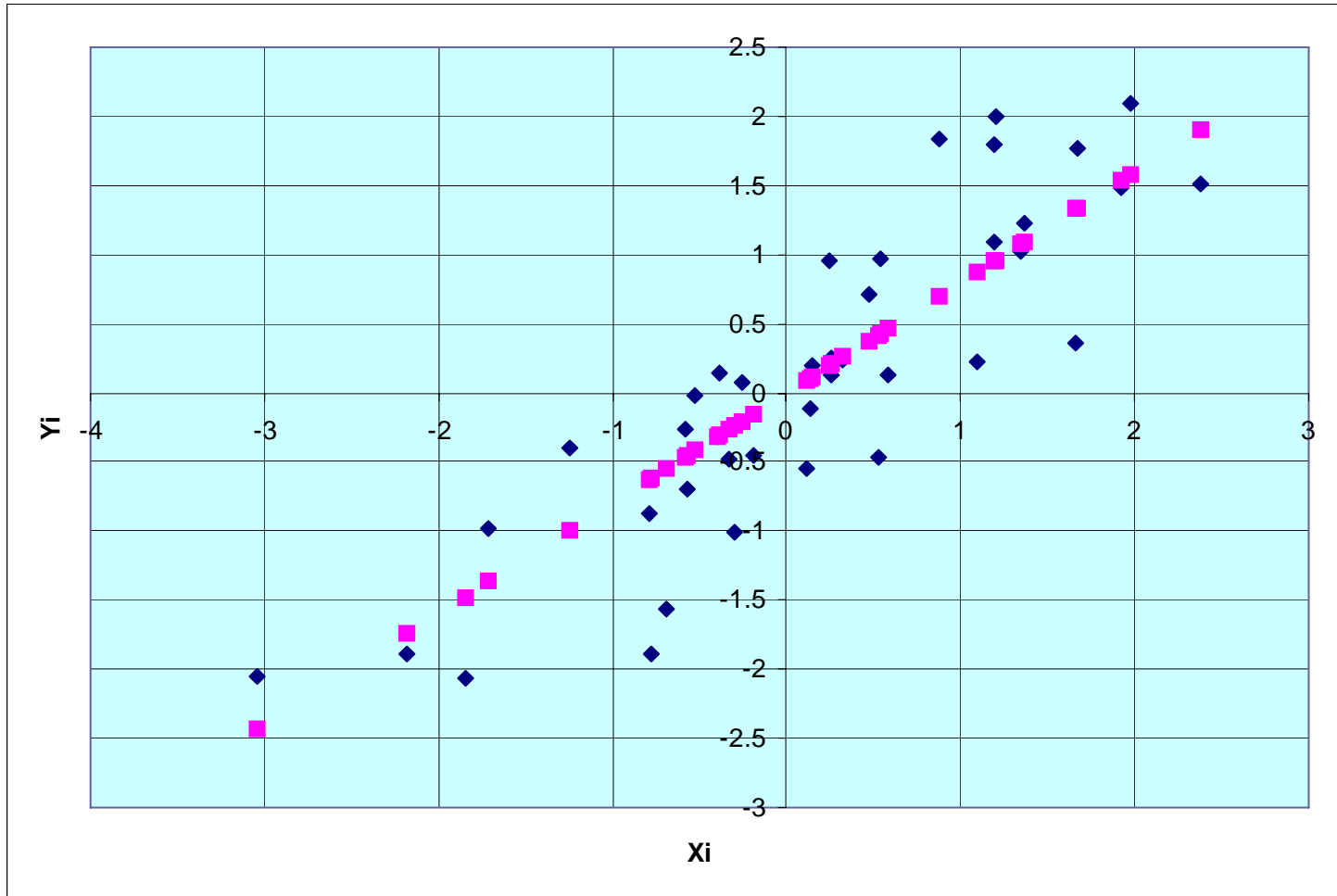
Regression: Straight line curve fitting

- Linear approximation: $\hat{y}_i = a + bx_i$
- Minimize residuals: $\frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2$
- Intercept: $a = \bar{y} - b \bar{x}$
- Slope: $b = r (s_y/s_x)$
- Regression: $\hat{y}_i = \bar{y} + r \frac{s_y}{s_x} (x_i - \bar{x})$
- Residual error: $s_y^2 (1 - r^2)$

Regression from data



Regression just for data points



Correlation for random variables

- X takes n values a_i with probability $P\{X=a_i\}$
- Y takes m values b_j with probability $P\{Y=b_j\}$
- Means: $\mu_X = E\{X\}$; $\mu_Y = E\{Y\}$
- Variances: $\text{Var}(X) = \sigma_X^2 = E\{(X-\mu_X)^2\}$;
 $\text{Var}(Y) = \sigma_Y^2 = E\{(Y-\mu_Y)^2\}$
- Correlation: $r_{XY}\sigma_X\sigma_Y = E\{[X - \mu_X][Y - \mu_Y]\} =$

$$\sum_{j=1}^m \sum_{i=1}^n (a_i - \mu_X)(b_j - \mu_Y)P\{X = a_i, Y = b_j\}$$

Regression for two random variables

- Linear fit: $\hat{Y} = a + b X$
- Mean-Squared-Error (MSE) = $E\{(\hat{Y} - Y)^2\}$
- Slope: $b = r_{XY} \sigma_Y / \sigma_X$
- Intercept: $a = \mu_Y - b \mu_X$
- $MSE = \sigma_Y^2(1 - r^2)$
- Regression: $\hat{Y} = \mu_Y + r_{XY} \frac{\sigma_Y}{\sigma_X} (X - \mu_X)$

Joint probabilities for two dice

Y=	2	3	4	5	6	7	8	9	10	11	12	P(X)
X =												1/6
1	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	0	1/6
2	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	1/6
3	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	1/6
4	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	1/6
5	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	1/6
6	0	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	1/6
P(Y)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	

Correlation for the two dice

- Means: $E\{X\} = 3.5$; $E\{Y\} = 7.0$

- Variances:

$$-\sigma_X^2 = \sum_{k=1}^6 (k - 3.5)^2 \frac{1}{6} = 35/12$$

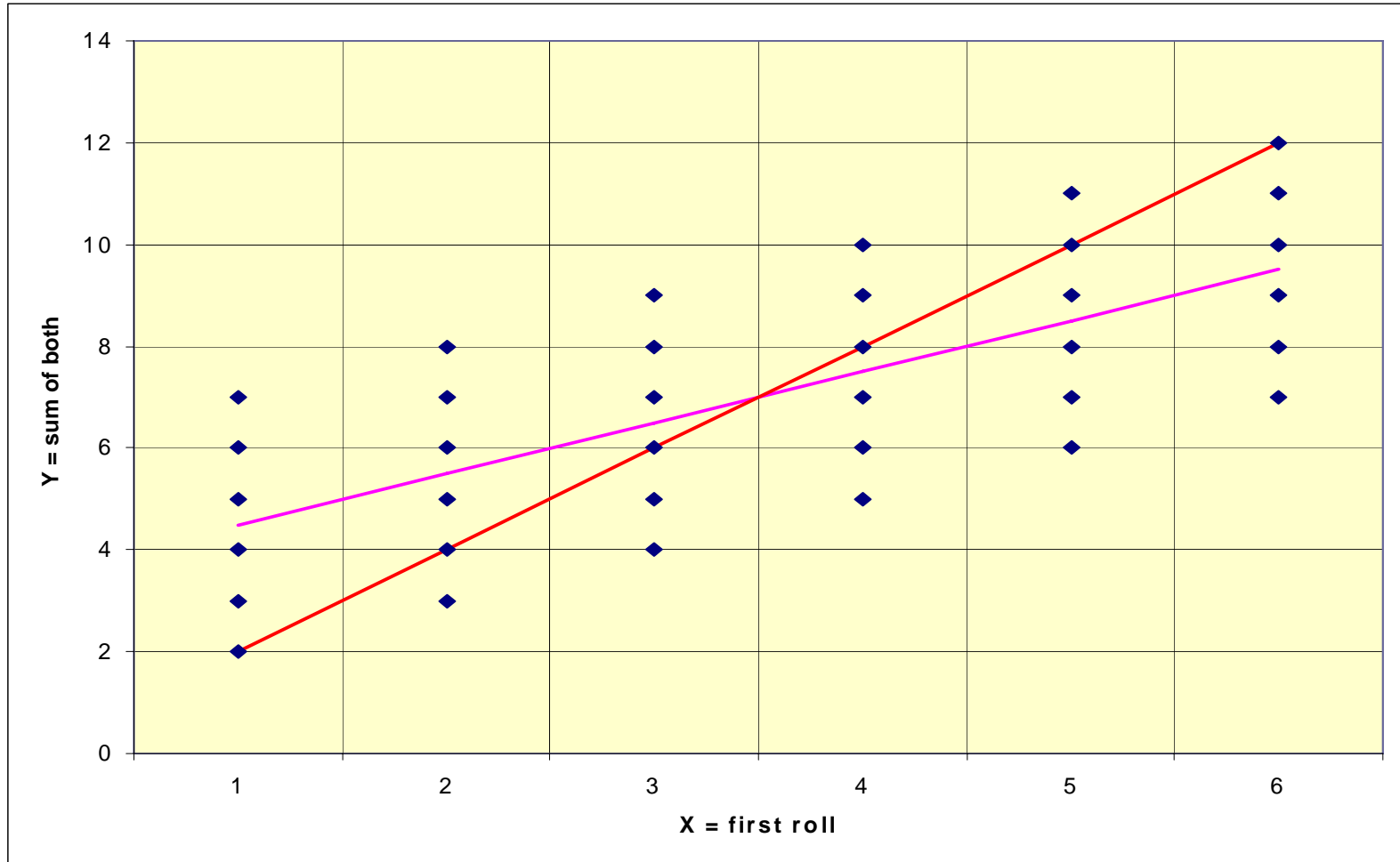
$$-\sigma_Y^2 = \sum_{k=2}^7 (k-7)^2 (k-1)/36 + \sum_{k=8}^{12} (k-7)^2 (13-k)/36 = 35/6$$

- $r_{XY} \sigma_X \sigma_Y = \sum_{\ell=2}^{12} \sum_{k=1}^6 (k-3.5)(\ell-7)P\{X=k, Y=\ell\} = 35/12$

Regressions for the dice

- $r_{XY} = 35/(12\sigma_X\sigma_Y) = 0.707$
- $\hat{Y} = \mu_Y + r_{XY} \frac{\sigma_Y}{\sigma_X} (X - \mu_X) = X + 3.5$
- $\hat{X} = \mu_X + r_{XY} \frac{\sigma_X}{\sigma_Y} (Y - \mu_Y) = Y/2$

The regression lines for the two dice



Summary of the two dice

- X = the value of the first roll
- Y = the sum of the two rolls
- Mean of X = 3.5; Mean of Y = 7
- $\text{Var}(X) = 35/12$; $\text{Var}(Y) = 35/6$
- Regression of Y from X = $X+3.5$
- Regression of X from Y = $Y/2$

Estimation of Signal in Noise

- X = the value of the signal (mean μ_X)
- N = the value of the noise (mean zero)
- $Y = X + N$ = the value of the observation
- Mean of $Y = \mu_Y = \mu_X$
- Variance of $Y = \text{Var}(Y) = \text{Var}(X) + \text{Var}(N)$
- Correlation = $r_{XY} = \sigma_X/\sigma_Y$
- Estimate of $X = \mu_X + (\sigma_X/\sigma_Y)^2(Y - \mu_Y)$
- Error Variance = $\text{Var}(X)/(1+\rho)$;
 - Signal-to Noise Ratio: $\rho = (\sigma_X/\sigma_N)^2$

Multiple Random Variables

- Mean of a sum:
 - $W = X + Y + Z$
 - Mean of W = Sum of the means of X, Y, Z
- Variance of a sum:
 - $\text{Var}(W) = \text{Var}(X) + \text{Var}(Y) + \text{Var}(Z) + 2r_{XY}\sigma_X\sigma_Y + 2r_{XZ}\sigma_X\sigma_Z + 2r_{YZ}\sigma_Y\sigma_Z$
 - $\text{Var}(\text{Sum}) = \text{Sum}(\text{Variances})$ for Uncorrelated Variables

The Law of Large Numbers

- X_i $i = 1, 2, \dots, n$, are uncorrelated random variables
 - Mean of $X_i = \mu$
 - Variance of $X_i = \sigma^2$
 - $Y = (X_1 + X_2 + \dots + X_n)/n$
 - Mean of $Y = \mu$
 - Variance of $Y = \sigma^2/n$
- As n becomes large, then Y tends to μ :
- If we add the variables and divide by n we approach the value of the mean if n is large

Central Limit Theorem

- X_i $i = 1, 2, \dots, n$, are independent random variables
 - Mean of $X_i = \mu$
 - Variance of $X_i = \sigma^2$
- $W = (X_1 + X_2 + \dots + X_n)$
 - Mean of $W = n\mu$
 - Variance of $W = n\sigma^2$
- $f_W(w) = [\exp\{-(w-n\mu)^2/(2n\sigma^2)\}]/[\sigma\sqrt{(2\pi n)}]$
 - Density of a sum of many independent variables approaches Normal

Example of a Channel

- Mean per user = 30kb/s
- Variance of each user = 400 (kb/s)²
- Mean of n users = 30n
- Variance of n users = 400n
- Standard deviation of n users = 20√n
- For capacity of 1,500 kb/s, and drop probability of 1% we can serve 40 users

Channel Example (cont.)

- If capacity is increased by 10 to 15,000
 - $n = 466$
- If capacity is increased by 100 to 150,000
 - $n = 4,891$
- This illustrates the effect of statistical multiplexing

Packet Switched Network

- Maximum transmission rate = R
- Probability of a user transmitting is $p=\alpha$
- Average transmission rate = $nR\alpha$
- Variance of transmission rate = $nR^2\alpha(1-\alpha)$
- $P_C = P\{\text{Congestion}\} = P\{\text{Total rate} > \text{Capacity}\}$
- $P_C = 1 - \Phi\left(\frac{(C/R) - n\alpha}{\sqrt{n\alpha(1-\alpha)}}\right)$

Example

- If $C=24R$ ($C=1.536\text{Mbps}$ and $R=64\text{kbps}$)
 - $P_c = 0.001$, then $n=133$
- If $C=240R$ ($C=15.36\text{Mbps}$ and $R=64\text{kbps}$)
 - $P_c = 0.001$, then $n=1986$
 - (49% more than the factor of 10)
- As C is increased we can approach the 10% utilization factor

Example 5.6

- $n=100$, $P\{\text{Success}\} = p = 0.3$
- W = number of successful messages
- $P\{20 < W < 40\}$ = approximately normal
 - Mean = $p = 0.3$
 - Variance = $p(1-p) = 0.21$
- $P\{20 < W < 40\} = \Phi(2.18) - \Phi(-2.18) = 0.9854 - 0.0146 = 0.9708 \approx 97\%$