

Random variables

A variable defined by the probabilities of each possible value in the population.

Probabilities show proportion of population with each possible value. This is how we move from yes/no probabilities to variables.

Probability distribution, $P(X)$ —the probability that each value will occur.

Continuous Random Variables—can take on any value in an interval of numbers, and in theory can be measured to any degree of accuracy.

Discrete Random Variable--Can take on only a finite countable set of values.

Marginal probabilities: Overall probabilities

Cumulative probabilities: Adding up probabilities for ranges of values.

Conditional probabilities: probabilities of values for subset of cases.

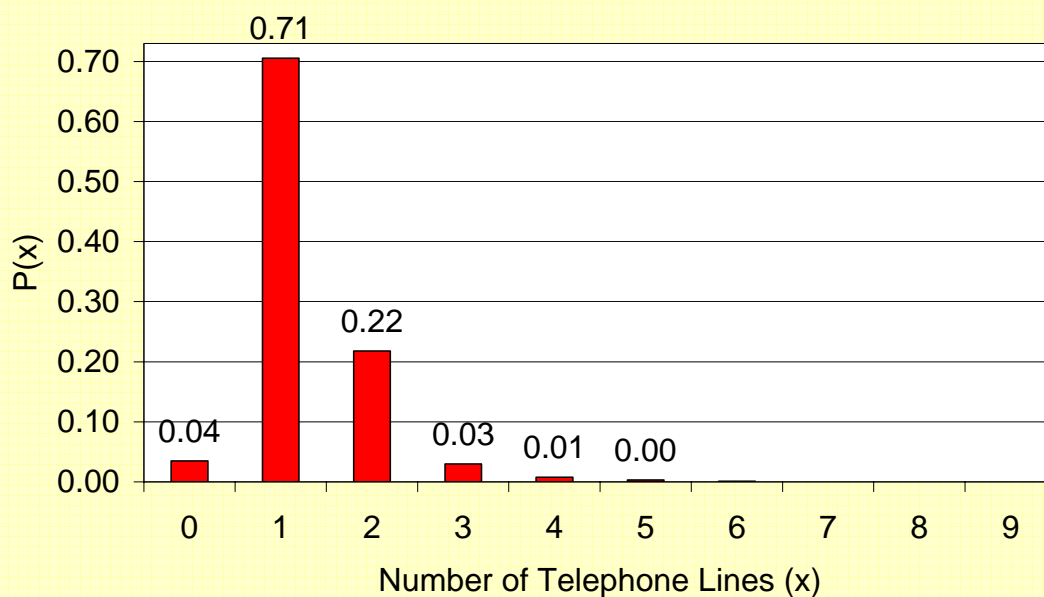
Joint probabilities: probabilities of combinations of values based on 2 or more variables.

The 2008 Washington State Population survey is a telephone survey of households.

Telephone surveys introduce some bias by excluding households that don't have telephones. *What do we know about the probability distribution of the number of phones in WA State households?*

<u>number of telephones,x</u>	<u>P(x)</u>
0	0.03500
1	0.70553
2	0.21769
3	0.02966
4	0.00775
5	0.00332
6	0.00088
7	0.00002
8	0.00000
9	0.00015
Total	1.00000

Probability Histogram of Telephone Coverage in Washington



[Data from 1990 Census for WA state.]

What is the probability that a household will have no telephone?

What is the probability that a household will have 1 or fewer telephone lines?

How about 4 or fewer?

What is the probability that a household will have 2 to 4 phone lines?

What is the probability a household will have 1 or fewer phone lines or more than 4 phone lines?

Means and Variances are variable values weighted by probabilities

Expected Value (population mean)

The sum of values weighted by the probabilities of those values.

$$\mu = E(X) = \sum xP(x)$$

Variance:

Sum of the squared deviations weighted by probabilities.

$$\sigma^2 = V(X) = E[(X - \mu)^2] = \sum (x - \mu)^2 P(x)$$

or

$$\sigma^2 = V(X) = E(X^2) - [E(X)]^2$$

The second equation is a shortcut—the mean of X-squared minus the squared mean of X.

Standard Deviation:

$$\sigma = \sqrt{\sigma^2} = SD(X)$$

The standard deviation is the square root of the variance.

Conditional Mean and SD—

- Mean for subset of cases
- Uses conditional probabilities to re-weight values

Functions of Random Variables—

What happens to mean and SD when values change (but probability weights stay the same?)

- Linear functions (adding, subtracting, multiplying or dividing all values by same number) : mean and SD change by same function
- nonlinear functions: no simple formula—need to recalculate mean and SD

What is the average number of telephones in Washington Households and how much does the number vary from the average?

# of Phones x	Frequency	P(x)	xP(x)	Approach 1: Variance			Approach 2: Variance	
				x-μ	(x- μ) ²	(x-μ) ² P(x)	x ²	x ² P(x)
0	198,286	0.04	0.00	-1.3	1.65	0.06	0	0.00
1	4,142,030	0.71	0.71	-0.3	0.08	0.06	1	0.71
2	1,278,026	0.22	0.44	0.7	0.51	0.11	4	0.87
3	174,110	0.03	0.09	1.7	2.94	0.09	9	0.27
4	45,499	0.01	0.03	2.7	7.38	0.06	16	0.12
5	19,473	0.00	0.02	3.7	13.81	0.05	25	0.08
6	5,170	0.00	0.01	4.7	22.24	0.02	36	0.03
7	118	0.00	0.00	5.7	32.67	0.00	49	0.00
8	-	0.00	0.00	6.7	45.10	0.00	64	0.00
9	897	0.00	0.00	7.7	59.53	0.01	81	0.01
Sum	5,863,609	1.00	μ=1.28	32.16		σ²=0.45		2.10

MEAN:

$$\mu = E(X) = \sum xP(x) = \mathbf{1.28}$$

VARIANCE:

Approach 1 above

Approach 2 continued: $\sigma^2 = E(x^2) - E(x)^2 = \mathbf{2.10} - (\mathbf{1.28})^2 = \mathbf{.45}$

STANDARD DEVIATION: $\sigma = \sqrt{\sigma^2} = \mathbf{0.67}$

How would the mean and SD be different if we looked only at households in our survey?

See "Characteristics of Households With and Without Telephones" for more info.