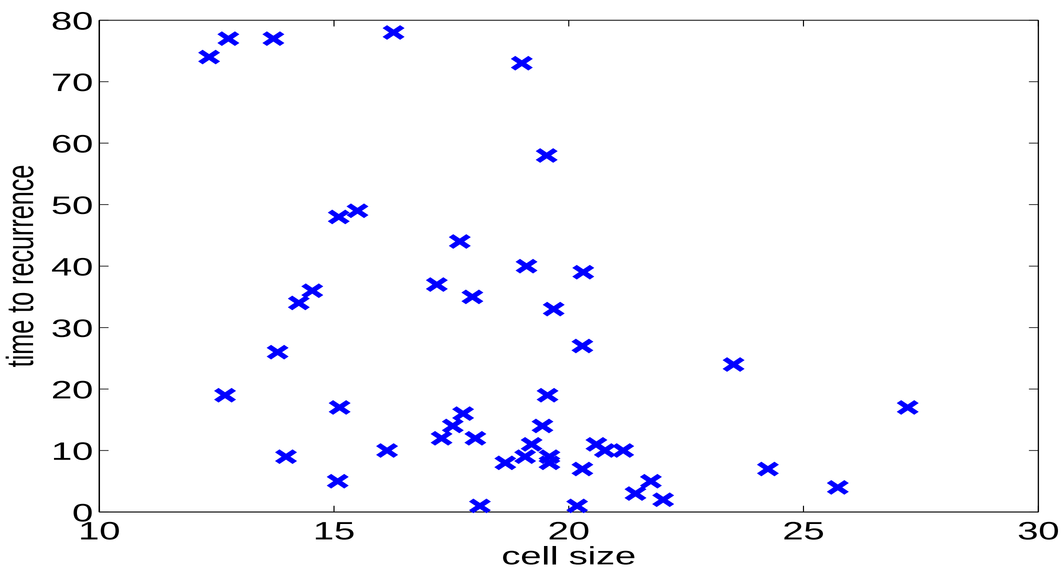
Testing the Statistical [In]Dependence of Random Variables

Examples

- Is there a relationship between tumor cell size and recurrence?

	recur	not recur
cell size > 17.4	31	16
cell size ≤ 17.4	66	85

- Is there a relationship between tumor cell size and time-to-recurrence?



Today

- Recall: Dependent and independent r.v.'s
- Are two discrete r.v.'s related?
  - One answer: The chi-square (χ^2) test.
- Are two continuous r.v.'s related?
  - Why the general problem is difficult.
  - Linear correlation.
  - Regression as a measure of relatedness.

Dependent and independent r.v.'s

- R.v.'s X and Y (discrete or continuous) are defined to be independent if, for all x and y,

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

	$X = 1$	$X = 2$	$X = 3$	$P(Y)$
$Y = A$	0.08	0.2	0.12	0.4
$Y = B$	0.12	0.3	0.18	0.6
$P(X)$	0.2	0.5	0.3	

- X and Y are dependent if, for some x and y,

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

	$X = 1$	$X = 2$	$X = 3$	$P(Y)$
$Y = A$	0.1	0.2	0.1	0.4
$Y = B$	0.1	0.3	0.2	0.6
$P(X)$	0.2	0.5	0.3	

In terms of conditional probability...

- Alternatively, X and Y are independent if for all x and y

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

because then $P(X, Y) = P(X|Y)P(Y) = P(X)P(Y)$.

- Intuitively, X and Y are independent if knowing Y tells you nothing about X. (I.e., doesn't help you predict X.)

- Same thing applies with X and Y reversed.

Example: independent r.v.'s

Joint:

	$X=1$	$X=2$	$X=3$	$P(Y)$
$Y=A$	0.08	0.2	0.12	0.4
$Y=B$	0.12	0.3	0.18	0.6
$P(X)$	0.2	0.5	0.3	

$P(X|Y)$:

	$X=1$	$X=2$	$X=3$
$Y=A$	0.2	0.5	0.3
$Y=B$	0.2	0.5	0.3

$P(Y|X)$:

	$X=1$	$X=2$	$X=3$
$Y=A$	0.4	0.4	0.4
$Y=B$	0.6	0.6	0.6

Example: dependent r.v.'s

Joint:

	$X=1$	$X=2$	$X=3$	$P(Y)$
$Y=A$	0.1	0.2	0.1	0.4
$Y=B$	0.1	0.3	0.2	0.6
$P(X)$	0.2	0.5	0.3	

$P(X|Y)$:

	$X=1$	$X=2$	$X=3$
$Y=A$	0.25	0.5	0.25
$Y=B$	0.166	0.5	0.333

$P(Y|X)$:

	$X=1$	$X=2$	$X=3$
$Y=A$	0.5	0.4	0.333
$Y=B$	0.5	0.6	0.666

Are two discrete r.v.'s related?

The χ^2 test: intuition

- Suppose X and Y are independent
- Suppose we observe N samples: (x_i, y_i).
- Let $N_{x,y}$ the number of observed pairs equal to (x, y).
- We expect $N_{x,y} \approx NP(x, y) = NP(x)P(y)$.

Data:

N=198	recur	not recur
cell size = big	31	16
cell size = small	66	85

Expected:

	recur	not recur	P(cell size)
cell size = big	23.3	24.2	0.24
cell size = small	73.7	76.7	0.76
P(recur)	0.49	0.51	

The χ^2 test: measuring discrepancy

- Let $\hat{P}(X)$ be the maximum likelihood estimate for $P(X)$, and likewise for Y.
- Let $E_{x,y} = NP(x)P(y)$ denote the expected number of observations of the pair (x, y).
- Compute $S = \sum_{x,y} \frac{(N_{x,y} - E_{x,y})^2}{E_{x,y}}$.
- If X and Y are truly independent, then S should be comparatively small.
- The larger S is, the greater is the discrepancy between the expectations and the observed data, and the greater the evidence that X and Y are dependent.

Example

case	$N_{x,y}$	$E_{x,y}$	$\frac{(N_{x,y}-E_{x,y})^2}{E_{x,y}}$
recur, cell size big	31	23.3	2.54
not recur, cell size big	16	24.2	2.77
recur, cell size small	66	73.7	0.80
not recur, cell size small	85	76.7	0.90

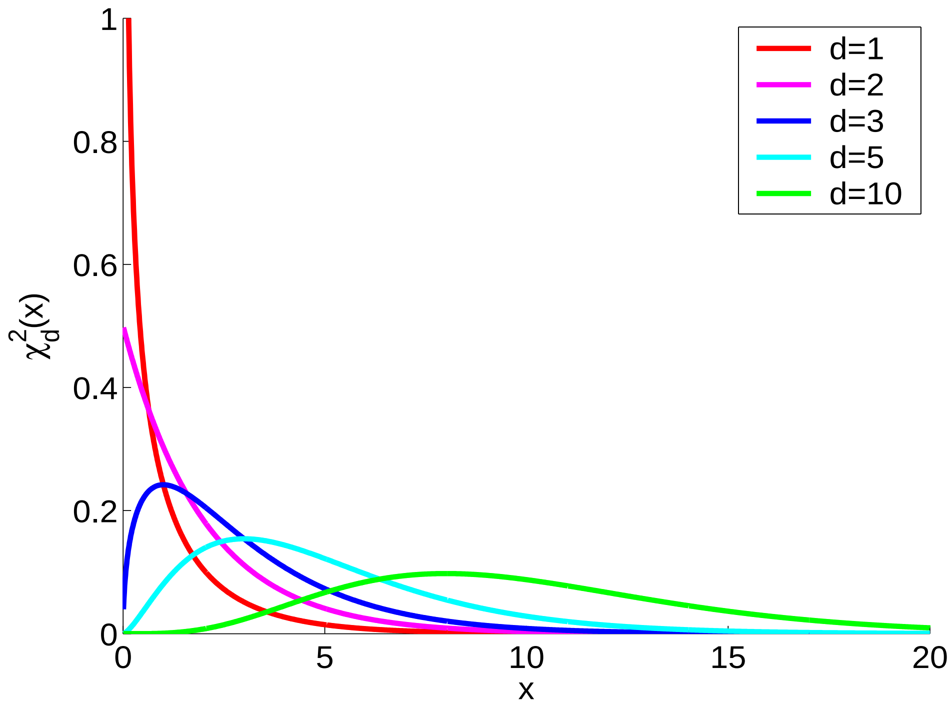
$$S = 7.03$$

- Is 7.03 big enough to claim the variables are related?

to be continued. . .

Aside: the χ^2 family of distributions

- χ^2_d is distributed as $Z_1^2 + Z_2^2 + \ldots + Z_d^2$, where each Z_i is a standard normal r.v. ($\mu = 0$,$\sigma = 1$)
- d is the "degrees-of-freedom"



Application to independence testing

- It turns out that, regardless of $P(X)$ and $P(Y)$, the value S computed in the χ^2 test is approximately distributed like $\chi^2_{(r-1)(c-1)}$ where
  - r is the number of different values Y can take. (The number of rows in the table.)
  - c is the number of different values X can take.
- (Hence, the name χ^2 test.)
- If S is unusually large for for a $\chi^2_{(r-1)(c-1)}$ random variable, this is taken as evidence for the dependence of X and Y.

Example continued

case	$N_{x,y}$	$E_{x,y}$	$\frac{(N_{x,y}-E_{x,y})^2}{E_{x,y}}$
recur, cell size big	31	23.3	2.54
not recur, cell size big	16	24.2	2.77
recur, cell size small	66	73.7	0.80
not recur, cell size small	85	76.7	0.90

$$S = 7.03$$

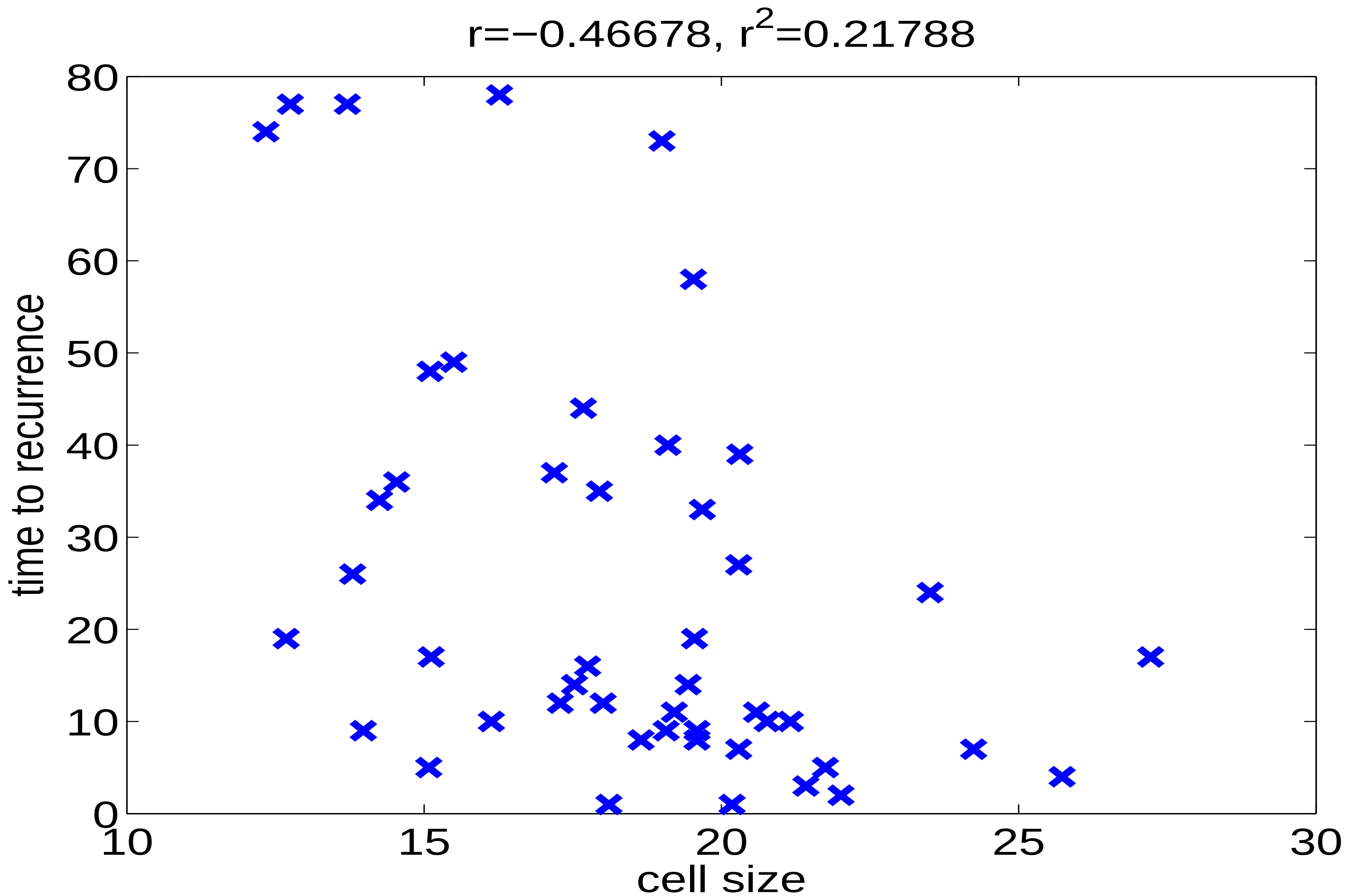
- Is 7.03 big enough to claim the variables are related?
- The probability that a χ_1^2 r.v. is ≥ 7.03 is less than 0.008, strong evidence of a dependence between X and Y.

Summary

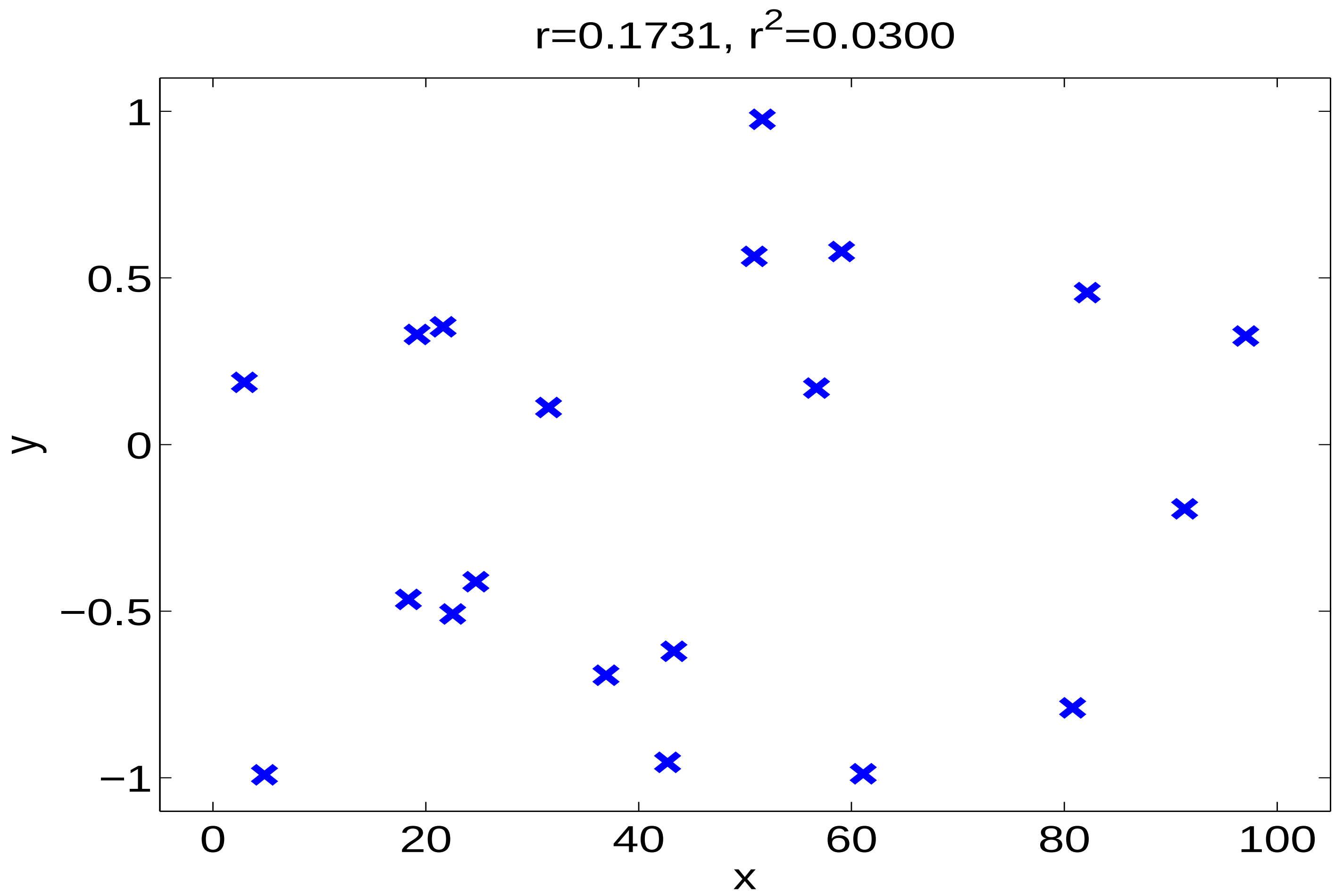
- The χ^2 test estimates whether or not there is a dependency between two discrete r.v.'s.
- The test is only approximate, and works best when the number of samples is large — particularly, when the number of samples in each cell is not too small. (≥ 5?)
- There are numerous variants of χ^2 as well as other tests for dependency between two discrete r.v.'s. (Such as Fisher's exact test.)

Are two continuous r.v.'s related?

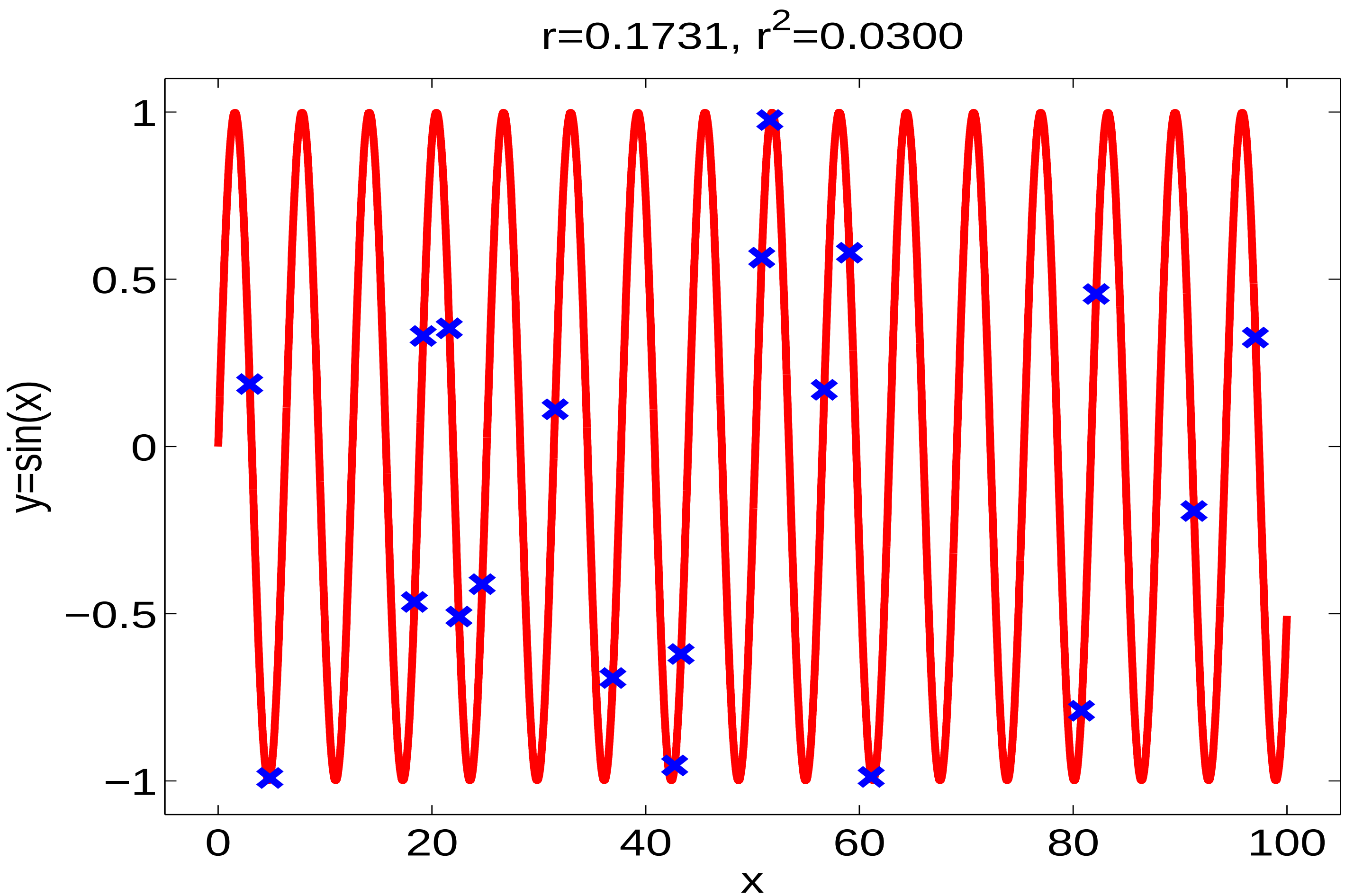
Cell-size versus Time-to-recurrence



Synthetic example



Synthetic example again



Relatedness of continuous r.v.'s

- The difficulty with testing for dependence of continuous r.v.'s is that their relationship can be arbitrarily complex.

- If we posit a specific kind of relationship, such a linear, *then* we can test how related the r.v.'s are—essentially by doing regression.

- If we can predict Y any better based on X than we can without X, then X and Y are dependent.

Linear correlation

- Given paired samples (x_i, y_i) distributed according to $P(X, Y)$, the [linear/Pearson's] correlation coefficient is

$$r = \sum_i \frac{(x_i - \mu_x)(y_i - \mu_y)}{\sigma_x \sigma_y}$$

where μ_x and μ_y are the sample means, and σ_x^2 and σ_y^2 are the sample variances.