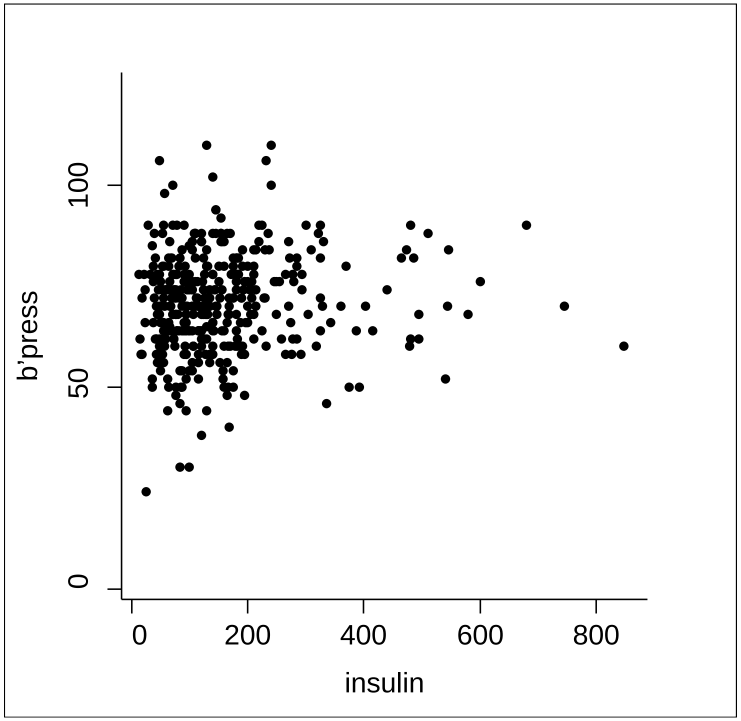
INFERENCE FOR DATA VISUALIZATION

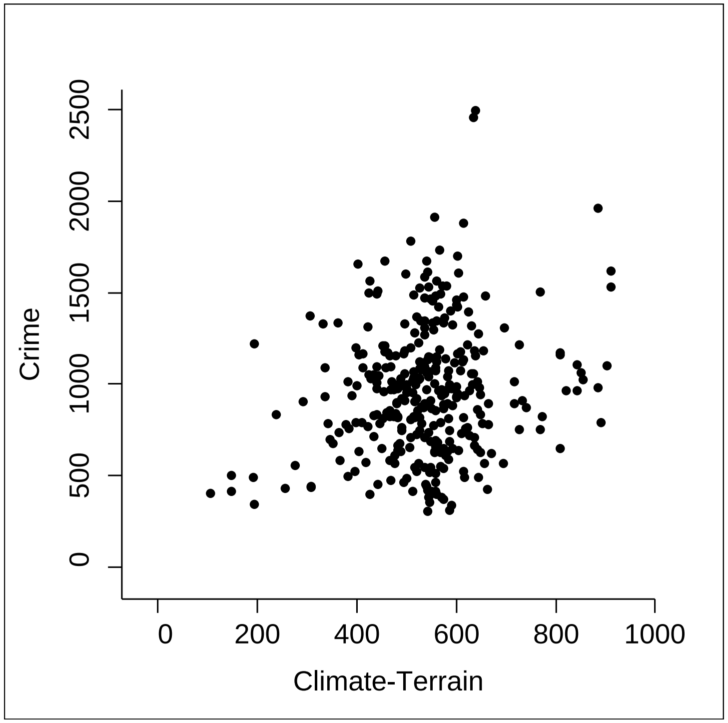
Andreas Buja

AT&T Labs

Shannon Laboratory, Florham Park, NJ

Two Examples:





Decreasing variance?

Positive correlation?

Some Questions:

- Is what we see REALLY there?
- What does it mean to be REALLY there?
- How prone is the eye to overinterpret?
- Is it true that looking at data invalidates inference?
- If inference for numbers is possible, why not for visual features?

What Does it Mean to be "Really There"?

An answer gleaned from statistical testing:

Under scenarios where the underlying feature is absent, the visible feature in the data is too unlikely to have arisen by chance.

- scenario where the feature is absent = null distribution
- underlying feature = a specific alternative
- visible feature = a statistical test

Visual Perception as a Statistical Test

Visual feature detector = test function $\phi(\text{data})$ such that

$\phi(\text{data}) = 1$ if a feature is detected,

$\phi(\text{data}) = 0$ if no feature is detected.

Q: What are the null hypothesis and alternative for ϕ?

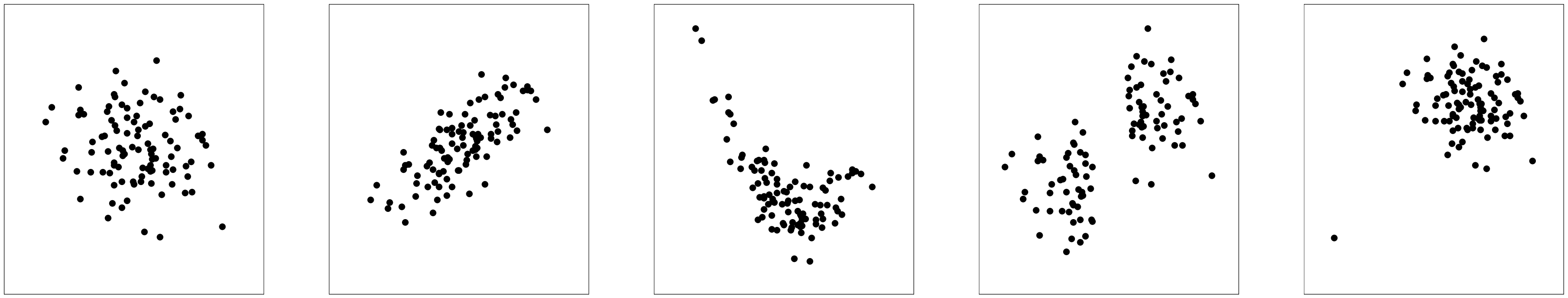
Given a visual test, what hypothesis and alternative?

Observation: In EDA, we don't know what feature we'll detect, so we have to include all of them.

$\Rightarrow$ *Interpretations:*

- Null hypothesis: "absence of **all** features" ($\forall$).
- Alternative: "presence of **some** feature" ($\exists$).

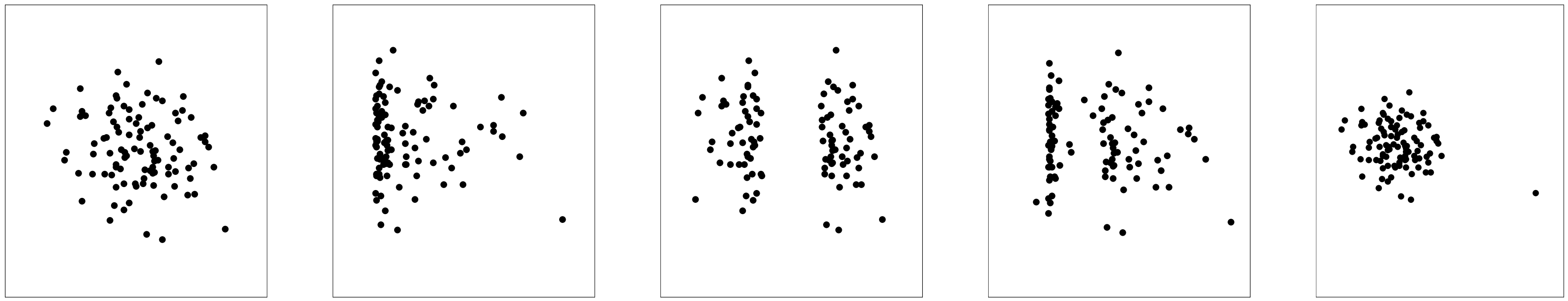
Example:



- We detect a linear increasing trend in an X-Y scatterplot.
- Had there been any other trend (nonlinear, decreasing, discontinuous,...), we would have detected it, too.
- In fact, we would have detected almost any type of dependence between X and Y...

$\Rightarrow$ The natural null hypothesis is independence of X and Y.

The Problem of Focusing Visual Detection



- If we're interested in dependence between X and Y, we must try to ignore marginal structure.

- The above plots differ only in the marginal structure of X; X and Y are independent.

$\Rightarrow$ It may be difficult to tailor visual detection to the structure of interest.

Significance Levels for Visual Detection

Recipe to establish a visual significance level:

1. If a null hypothesis can be simulated, create a large number (N-1) of views of simulated null data.
2. Randomly insert the view of the actual data $\Rightarrow$ N views.
3. Ask an uninvolved person to select the most special looking view.
4. If the selected view shows the actual data, the existence of a feature is significant at the level α=1/N.

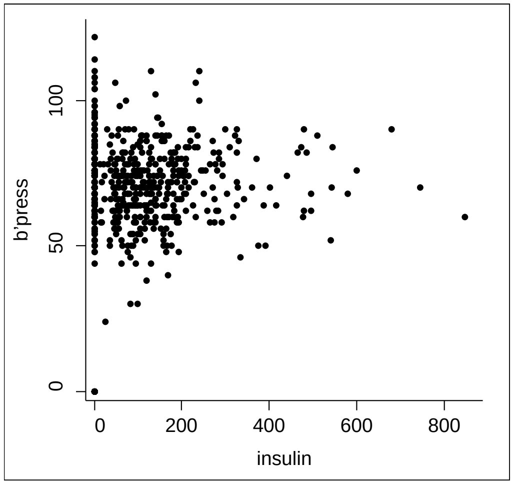
Examples of Null Hypotheses that Can be Simulated

- Any univariate distributional assumption, e.g., normality.
- Independence assumptions between two variables: shuffle X-values against Y-values, as in a permutation test.
- Exact tests, null hypotheses with Neyman structure: simulate the conditional distribution given the sufficient statistic.

The Pima Indian Diabetes Data

- 768 Pima Indians
- 2 of 8 variables:

  blood pressure vs. serum insulin
- From UC Irvine ML database



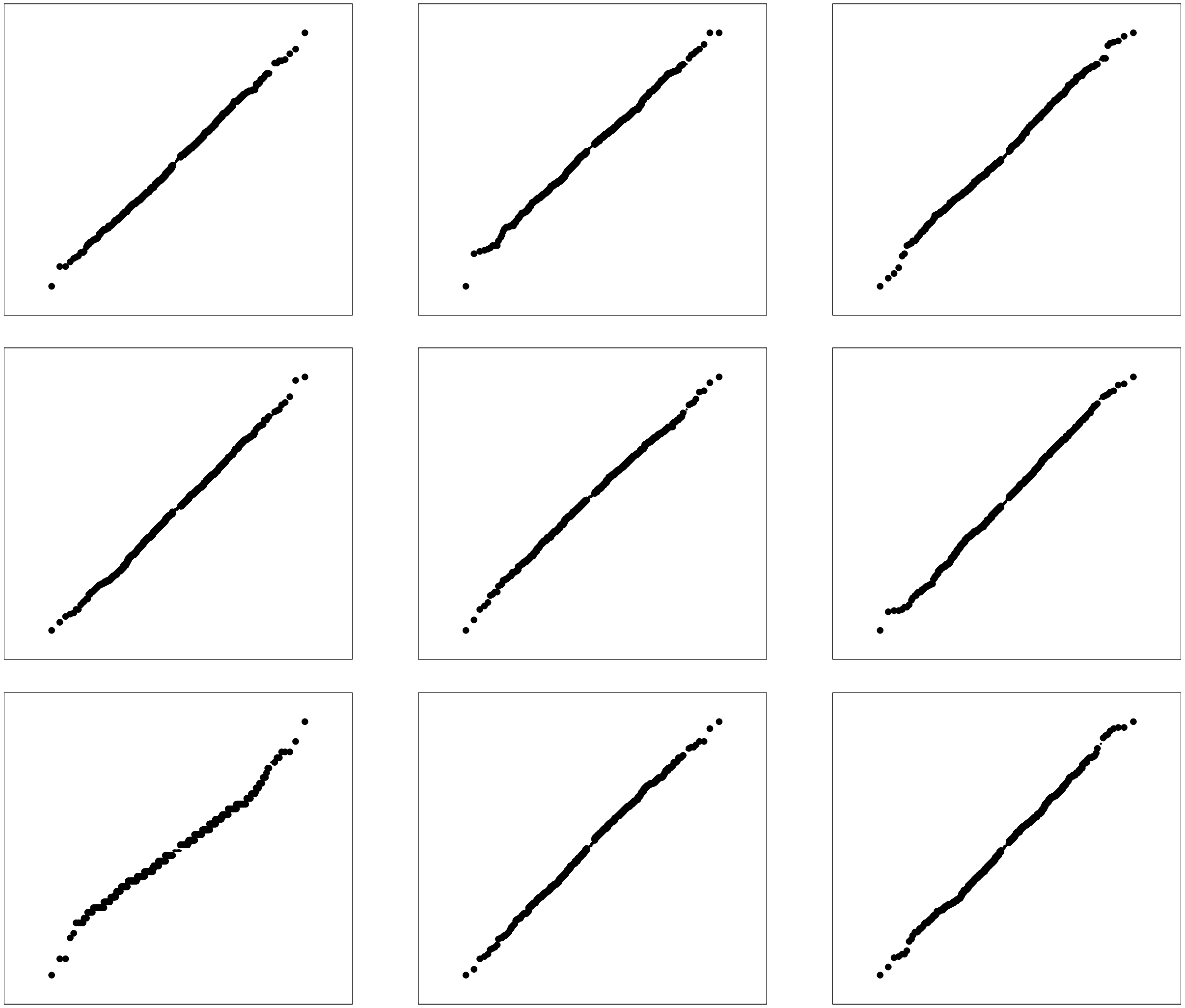
Example: A Visual Normality Test of the Pima Data

Which

one

is

b'press?



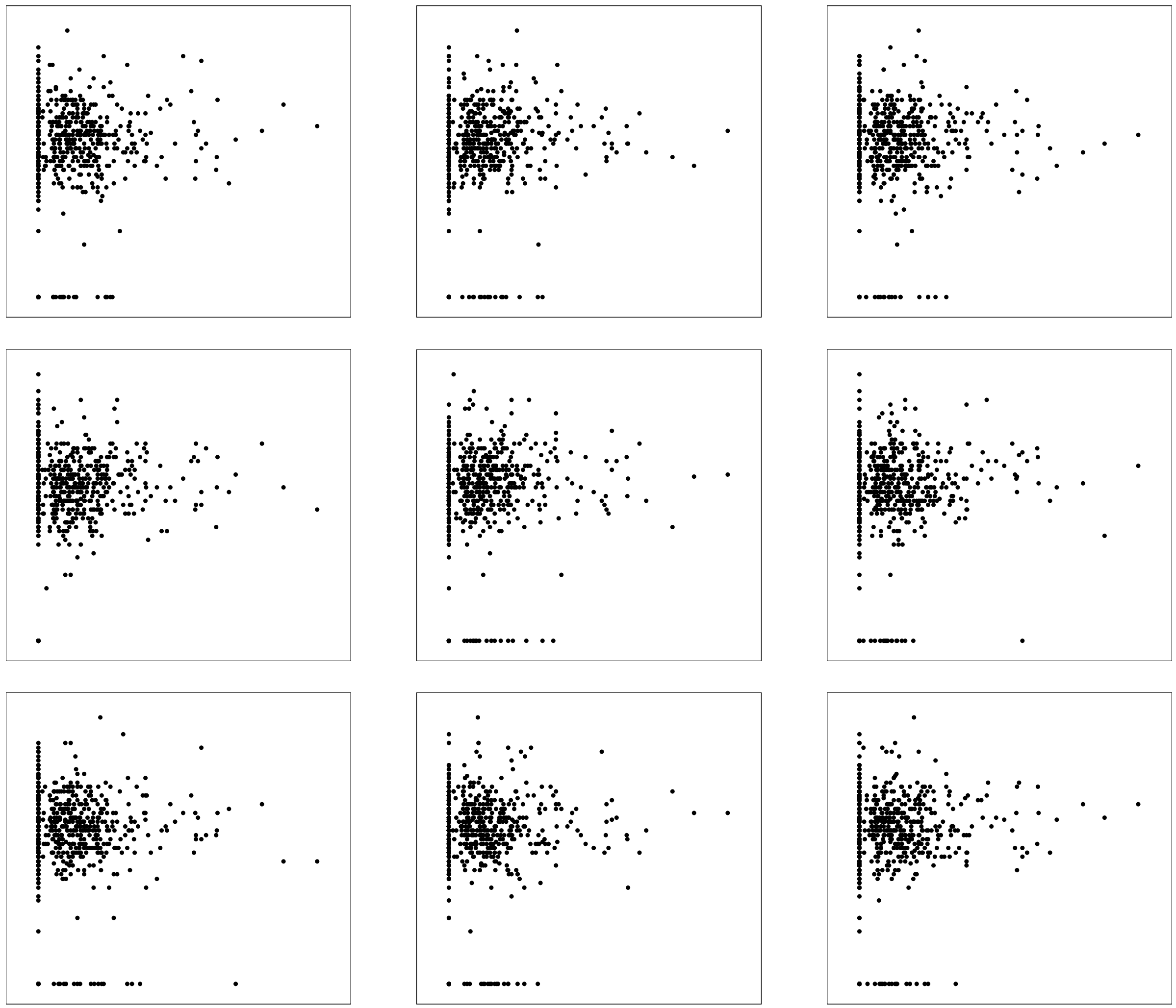
Example: A Visual Permutation Test of the Pima Data

Which

one is

b'press

vs insln?



Objections

- Is snooping the null data kosher?
  $\triangleleft$ Sure, as kosher as evaluating a test statistic on null data in a permutation test.

- Isn't inference invalidated by comparing the real data with null data?
  $\triangleleft$ True. But we're honest as long as we snoop on the null data AND the real data WITHOUT KNOWING A PRIORI which is the real data. [We may need an uninvolved judge.]

Objections (cont.)

- Aren't we unable to visually assess the whole course of our data explorations?
  $\triangleleft$ True; the opportunistic application of tests when snooping weakens their validity. But presence/absence of features usually needs no testing; tests are needed when in doubt.

- Don't we tailor the test to the feature we found by snooping?
  $\triangleleft$ If we do, it weakens the validity of the test. But if features concern general dependencies among variables, permutation tests of independence are broadly valid and not much tailored.

Conclusions

- Visual inference is often possible in principle.
- The human eye acts is a broad feature detector and general statistical test.
- For valid visual inference, it may be necessary to obey a mild testing regime:
  - Limit yourself to distributional assumptions and general dependencies to avoid tailoring of the null hypothesis...
  - Generate a large number of null pictures...
  - Use an uninvolved judge who is not acquainted with the data to avoid discrimination of the real data from null data due to prior knowledge...