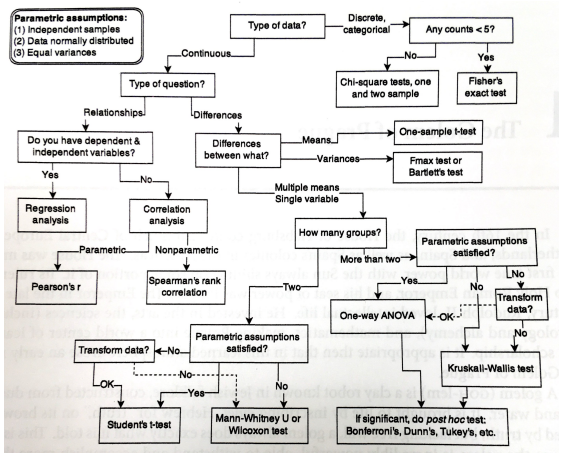
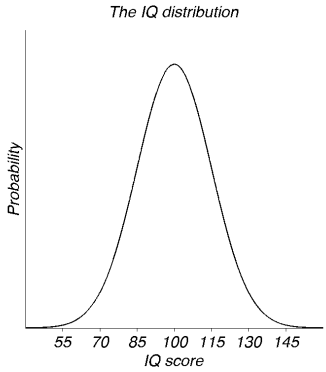
Classical “frequentist” statistical tests



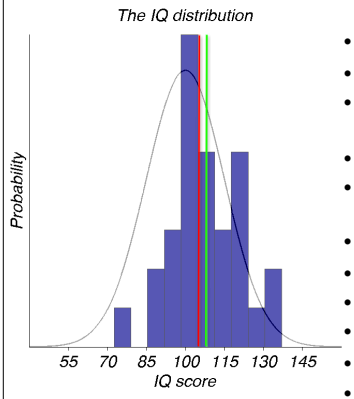
Statistical Rethinking, Richard McElreath

Classical/frequentist approach - *z*

- H_1: NZT improves IQ
- Null: H_0: it does nothing
- In the general population, IQ is known to be distributed normally with
- $\mu = 100$
- $\sigma = 15$
- We give the drug to 30 people and test their IQ.

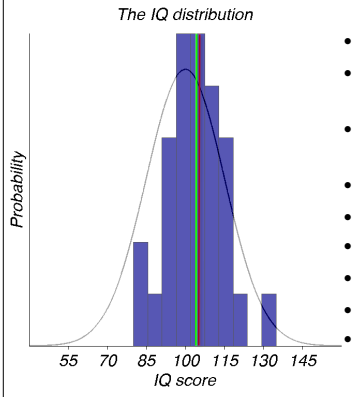


The *z*-test



- $\mu = 100$ (Population mean)
- $\sigma = 15$ (Population standard deviation)
- $N = 30$ (Sample contains scores from 30 participants)
- $x = 108.3$ (Sample mean)
- $\bar{z} = (x - \mu)/\text{SE} = (108.3\text{-}100)/\text{SE}$ (Standardized score)
- $\text{SE} = \sigma / \sqrt{N} = 15/\sqrt{30} = 2.74$
- Error bar/CI: ±2 SE
- $z = 8.3/2.74 = 3.03$
- $p = 0.0012$
- Significant?
- One- vs. two-tailed test

What if the measured effect of NZT had been half that?



- $\mu = 100$ (Population mean)
- $\sigma = 15$ (Population standard deviation)
- $N = 30$ (Sample contains scores from 30 participants)
- $\bar{x} = 104.2$ (Sample mean)
- $z = (\bar{x} - \mu)/\text{SE} = (104.2\text{-}100)/\text{SE}$
- $\text{SE} = \sigma / \sqrt{N} = 15/\sqrt{30} = 2.74$
- $z = 4.2/2.74 = 1.53$
- $p = 0.061$
- Significant?

Significance levels

- Are denoted by the Greek letter α.
- In principle, we can pick anything that we consider unlikely.
- In practice, the consensus is that a level of 0.05 or 1 in 20 is considered as unlikely enough to reject H_0 and accept the alternative.
- A level of 0.01 or 1 in 100 is considered "highly significant" or really unlikely.

Does NZT improve IQ scores or not?

Significant?	Reality: Yes	Reality: No
Yes	Correct	Type I error α-error False alarm
No	Type II error β-error Miss	Correct

Test statistic

- We calculate how far the observed value of the sample average is away from its expected value.
- In units of standard error.
- In this case, the test statistic is

$$z = \frac{\bar{x} - \mu}{SE} = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}}$$

- Compare to a distribution, in this case z or $N(0,1)$

Common misconceptions

Is "Statistically significant" a synonym for:
- Substantial
- Important
- Big
- Real

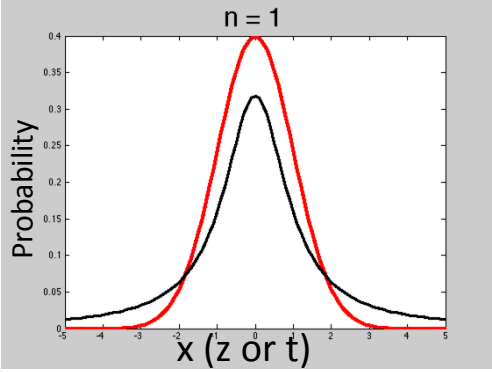
Does statistical significance gives the
- probability that the null hypothesis is true
- probability that the null hypothesis is false
- probability that the alternative hypothesis is true
- probability that the alternative hypothesis is false

Meaning of p-value. Meaning of CI.

Student's *t*-test

- σ not assumed known
- Use $$s^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}$$
- Why N-1? s is unbiased (unlike ML version), i.e., $E(s^2) = \sigma^2$
- Test statistic is $$t = \frac{\bar{x} - \mu_0}{s / \sqrt{N}}$$
- Compare to t distribution for CIs and NHST
- "Degrees of freedom" reduced by 1 to N-1

The *t* distribution approaches the normal distribution for large *N*



The *z*-test for binomial data

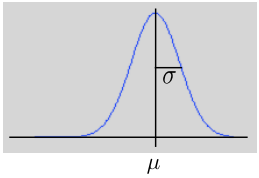
- Is the coin fair?
- Lean on central limit theorem
- Sample is *n* heads out of *m* tosses
- Sample mean: $\hat{p} = n / m$
- H_0: $p = 0.5$
- Binomial variability (one toss): $\sigma = \sqrt{pq}$, where $q = 1 - p$
- Test statistic: $$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / m}}$$
- Compare to *z* (standard normal)
- For CI, use $$\pm z_{\alpha/2} \sqrt{\hat{p}\hat{q} / m}$$

Many varieties of frequentist univariate tests

- χ^2 goodness of fit
- χ^2 test of independence
- test a variance using χ^2
- *F* to compare variances (as a ratio)
- Nonparametric tests (e.g., sign, rank-order, etc.)

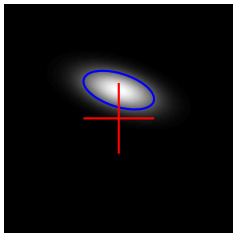
The Gaussian

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- parameterized by mean and stdev (position / width)
- joint density of two indep Gaussian RVs is circular! [easy]
- product of two Gaussian dists is Gaussian! [easy]
- conditionals of a Gaussian are Gaussian! [easy]
- sum of Gaussian RVs is Gaussian! [moderate]
- all marginals of a Gaussian are Gaussian! [moderate]
- central limit theorem: sum of many RVs is Gaussian! [hard]
- most random (max entropy) density with this variance! [moderate]

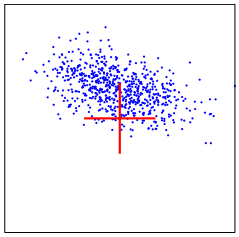
true density



Measurement (sampling)

Inference

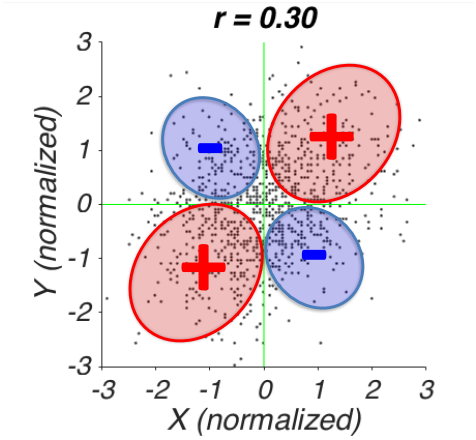
700 samples



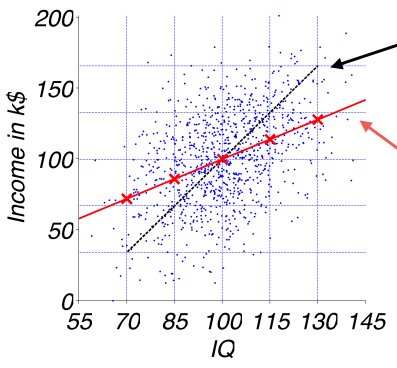
true mean: [0 0.8]
true cov: [1.0 -0.25
-0.25 0.3]

sample mean: [-0.05 0.83]
sample cov: [0.95 -0.23
-0.23 0.29]

Correlation: summary of data cloud shape



Correlation and regression

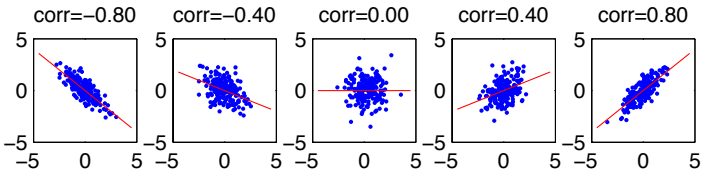


TLS (largest eigenvector)

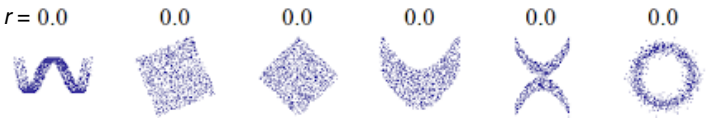
Least-squares regression

"Regression to the mean"

Correlation and regression

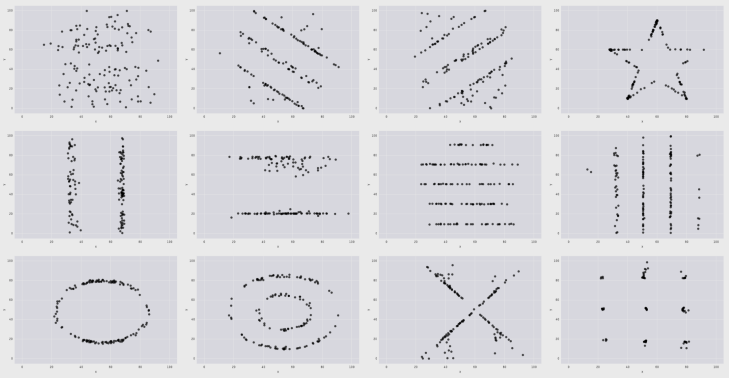


Independence implies uncorrelated,
but uncorrelated doesn't imply independent!



More extreme examples !

X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06

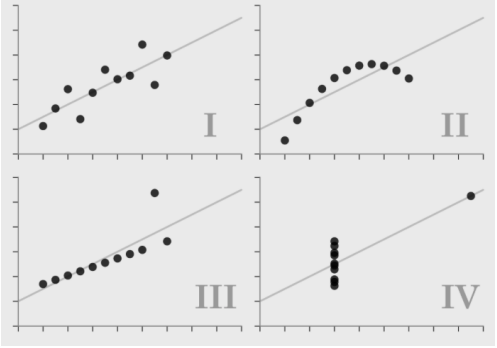


https://www.autodeskresearch.com/publications/samestats

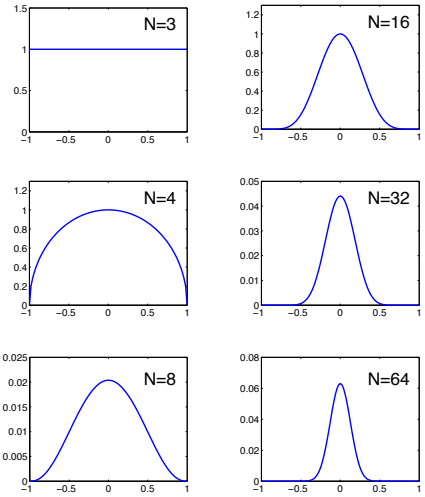
Correlation between variables does not explain their relationship

Anscombe's Quartet

Each dataset has the same summary statistics (mean, standard deviation, correlation), and the datasets are *clearly different*, and *visually distinct*.



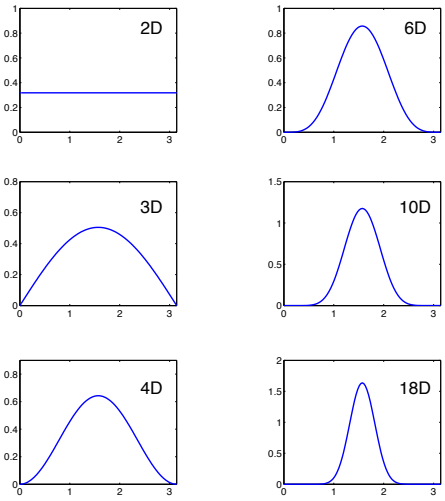
Correlation in N dimensions



Null Hypothesis: Distribution of normalized dot product of pairs of Gaussian vectors in N dimensions:

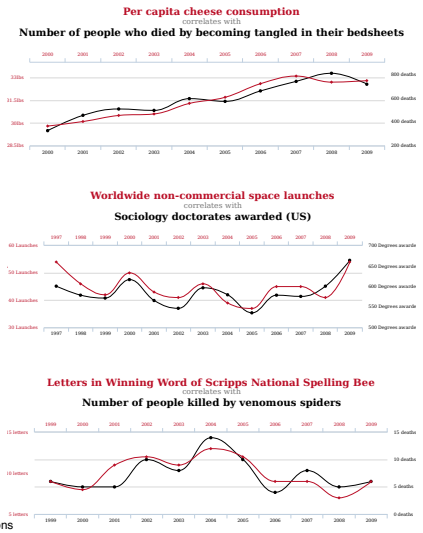
$$(1-d^2)^{\frac{N-3}{2}}$$

Distribution of angles of pairs of Gaussian vectors



sin(theta)^(N-2)

Correlation does not imply causation



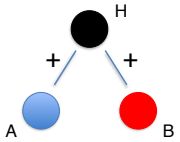
http://www.tylervigen.com/spurious-correlations

Correlation does not imply causation

- Beware selection bias
- Correlation does not provide a *direction* for causality. For that, you need additional (temporal) information.
- More generally, correlations are often a result of hidden (unmeasured, uncontrolled) variables…

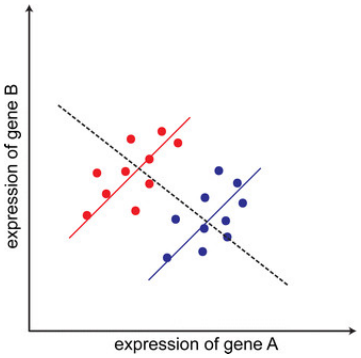
Example: conditional independence:

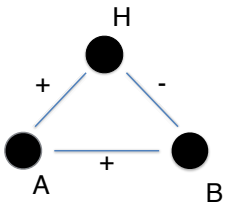
$$p(A,B \mid H) = p(A \mid H)\, p(B \mid H)$$



[on board: In Gaussian case, connections are explicit in the Precision Matrix]

Another example: Simpson's paradox

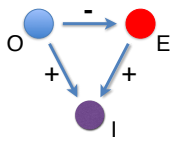




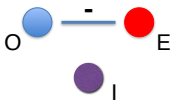
Milton Friedman's Thermostat

O = outside temperature (assumed cold)
I = inside temperature (ideally, constant)
E = energy used for heating

True interactions:



Statistical interactions, $P=C^{-1}$:



Statistical observations:

- O and I uncorrelated
- I and E uncorrelated
- O and E anti-correlated

Some nonsensical conclusions:

- O and E have no effect on I, so shut off heater to save money!
- I is irrelevant, and can be ignored. Increases in E cause decreases in O.

Statistical summary cannot replace scientific reasoning/experiments!

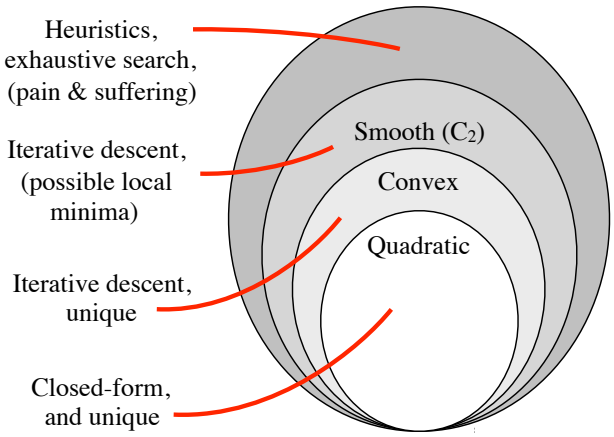
Summary: misinterpretations of Correlation

- Correlation => dependency, but non-correlation does not imply independence
- Correlation does not imply data lie on a line (subspace), with noise perturbations
- Correlation does not imply causation (temporally, or by direct influence)
- Correlation is only a descriptive statistic, and cannot replace the need for scientific reasoning/experiment

Taxonomy of model-fitting errors

- Optimization failures (e.g., local minima) [prefer convex objective, test with simulations]
- Overfitting [use cross-validation to select complexity, or to control regularization]
- Experimental variability (due to finite noisy measurements) [use math/distributional assumptions, or simulations, or bootstrapping]
- Model failures

Optimization…



Bootstrapping

- "The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps"
  [Adventures of Baron von Munchausen, by Rudolph Erich Raspe]
- A **(re)sampling** method for computing estimator distribution (incl. stdev error bars or confidence intervals)
- Idea: instead of running experiment multiple times, resample (with replacement) from the *existing* data. Compute an estimate from each of these "bootstrapped" data sets.

HEART ATTACK RISK FOUND TO BE CUT BY TAKING ASPIRIN

LIFESAVING EFFECTS SEEN

Study Finds Benefit of Tablet Every Other Day Is Much Greater Than Expected

[New York Times, 27 Jan 1987]

The summary statistics in the newspaper article are very simple:

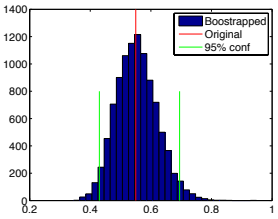
	heart attacks (fatal plus non-fatal)	subjects
aspirin group:	104	11037
placebo group:	189	11034

$$\widehat{\theta} = \frac{104/11037}{189/11034} = .55. \qquad (1.1)$$

If this study can be believed, and its solid design makes it very believable, the aspirin-takers only have 55% as many heart attacks as placebo-takers.

Of course we are not really interested in $\widehat{\theta}$, the estimated ratio. What we would like to know is θ, the true ratio

Histogram of bootstrap estimates:



=> with 95% confidence,

$$0.43 < \theta < 0.7$$

[Efron & Tibshirani '98]

	strokes	subjects	
aspirin group:	119	11037	
placebo group:	98	11034	(1.3)

For strokes, the ratio of rates is

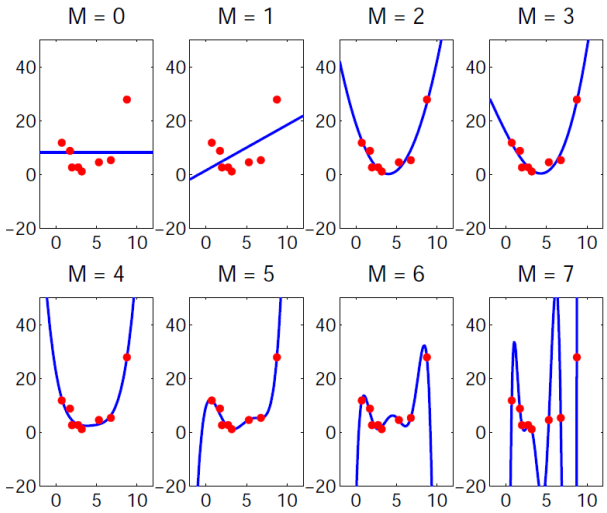
$$\widehat{\theta} = \frac{119/11037}{98/11034} = 1.21. \qquad (1.4)$$

It now looks like taking aspirin is actually harmful. However the interval for the true stroke ratio θ turns out to be

$$.93 < \theta < 1.59 \qquad (1.5)$$

with 95% confidence. This includes the neutral value $\theta = 1$, at which aspirin would be no better or worse than placebo vis-à-vis strokes. In the language of statistical hypothesis testing, aspirin was found to be significantly beneficial for preventing heart attacks, but not significantly harmful for causing strokes.

[Efron & Tibshirani '98]



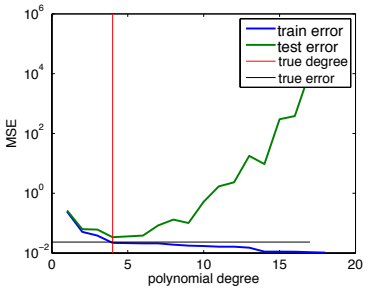
Cross-validation

A resampling method for constraining a model. Widely used to identify/avoid over-fitting.

(1) Randomly partition data into a "training" set, and a "test" set.
(2) Fit model to training set. Measure error on test set.
(3) Repeat (many times)

(4) Choose model that minimizes the cross-validated (test) error

Using cross-validation to select the degree of a polynomial model:

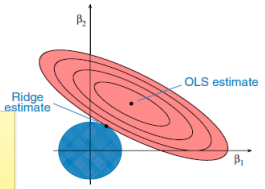


Ridge regression

(a.k.a. Tikhonov regularization)

Ordinary least squares regression:

$$\arg\min_{\vec{\beta}} ||\vec{y} - X\vec{\beta}||^2$$



Fix notation OLS, Ridge.
Redo figure: align ellipse with axes

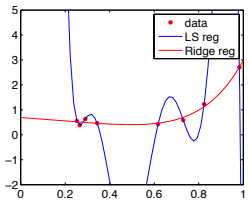
"Regularized" least squares

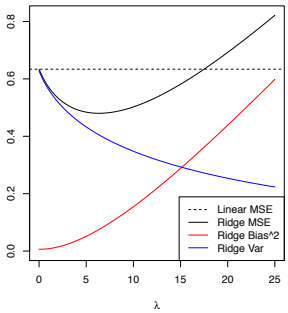
$$\arg\min_{\vec{\beta}} ||\vec{y} - X($$

Equivalent formulation: negative log posterior, assuming Gaussian likelihood & prior

Choose lambda by cross-validation

7th-order polynomial regression:





Linear regression:
Squared bias ≈ 0.006
Variance ≈ 0.627
Pred. error $\approx 1 + 0.006 + 0.627$
≈ 1.633

Ridge regression, at its best:
Squared bias ≈ 0.077
Variance ≈ 0.403
Pred. error $\approx 1 + 0.077 + 0.403$
≈ 1.48

from http://www.stat.cmu.edu/~ryantibs/datamining/

L_1 regularization

(a.k.a. least absolute shrinkage and selection operator - LASSO)

$$\arg\min_{\vec{\beta}} ||\vec{y} - X\vec{\beta}||^2 + \lambda \sum_k |\beta_k|$$

L_1 norm (still convex)

modify figure: align ellipse with axes, to show that some betas are turned off

Using an absolute error[illegible]erm promotes binary *selection* of regr