Pearson's chi-squared test

- The Pearson chi-squared test provides us with a way to test whether observed count data differs from some specific expected values that define the null hypothesis.

- The chi-squared statistic is a measure of the goodnessof-fit of the data to the model

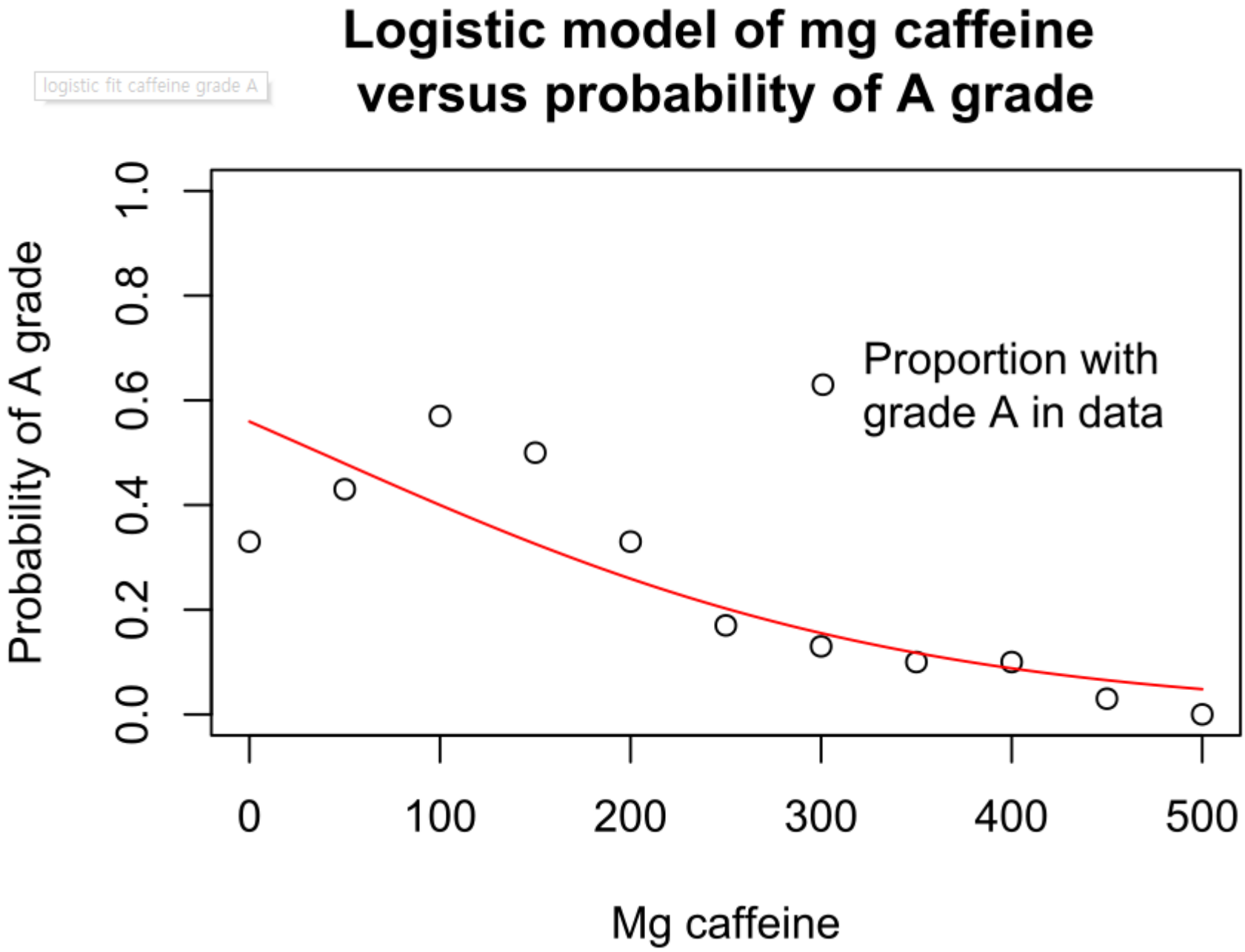
(Mendel's peas)

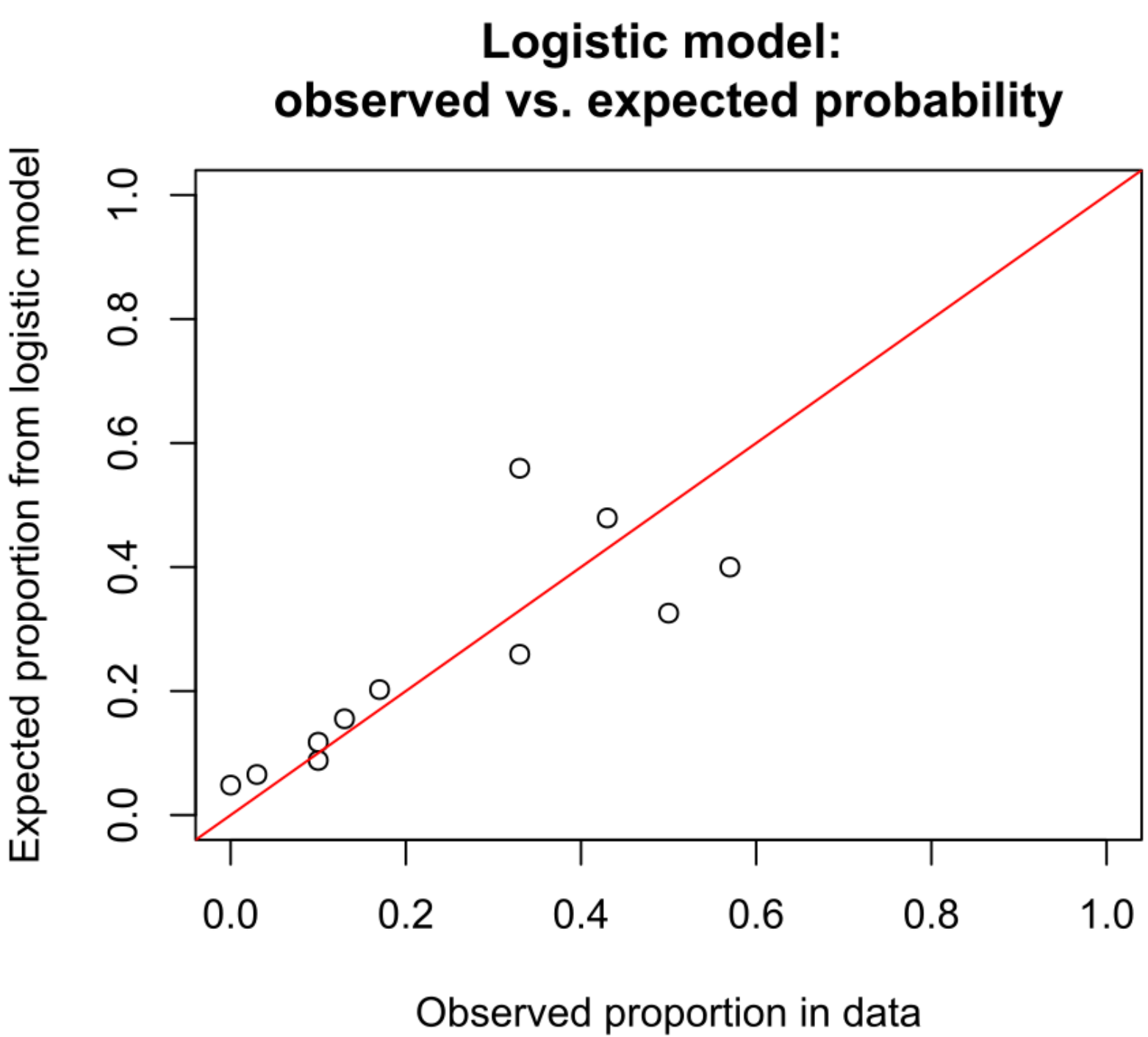
- Mendel bred peas with round yellow seeds and wrinkled green seeds. There are four types of progeny: round yellow, wrinkled yellow, round green, and wrinkled green. The number of each type is multinomial with probability p = (p1, p2, p3, p4). His theory of inheritance predicts that p is equal to (9/16, 3/16, 3/16, 1/16)

Hosmer–Lemeshow test

- a statistical test for goodness of fit for logistic regression models

logistic fit caffeine grade A



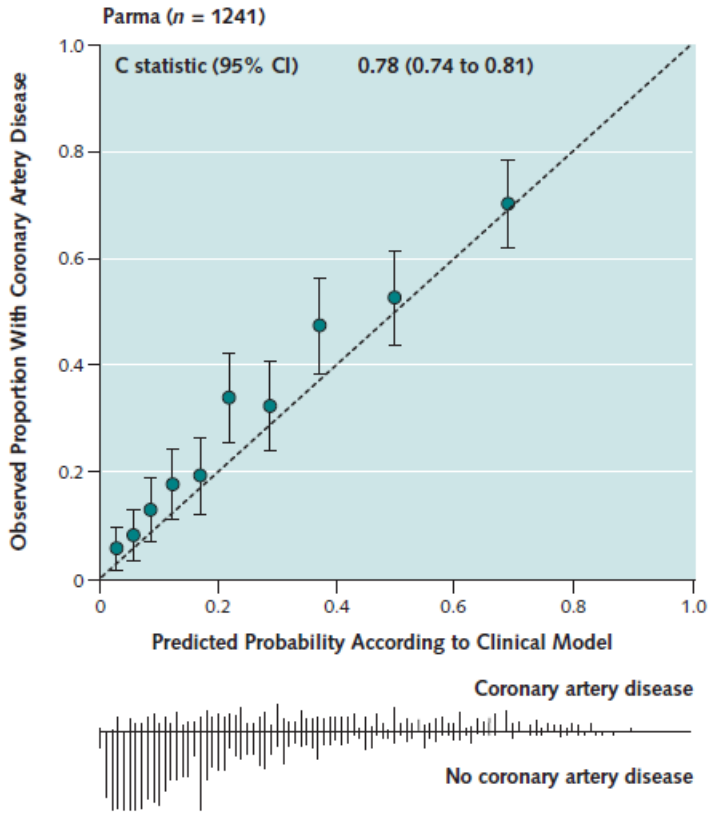

Logistic model:
observed vs. expected probability
Expected proportion from logistic model
0.0
0.2
0.4
0.6
0.8
1.0
Observed proportion in data

Model performance

- **Calibration**
  - the **agreement** between outcome predictions from the model and the observed outcomes.
  - a model is said to be well calibrated if, for every group of, say, 100 individuals, each with a mean predicted risk of x%, close to x indeed have the outcome.
  - Hosmer–Lemeshow test
- **Discrimination**
  - to the ability of a prediction model to **differentiate** between those who do or do not experience the outcome event
  - Concordance index (c-index)

Model performance - calibration

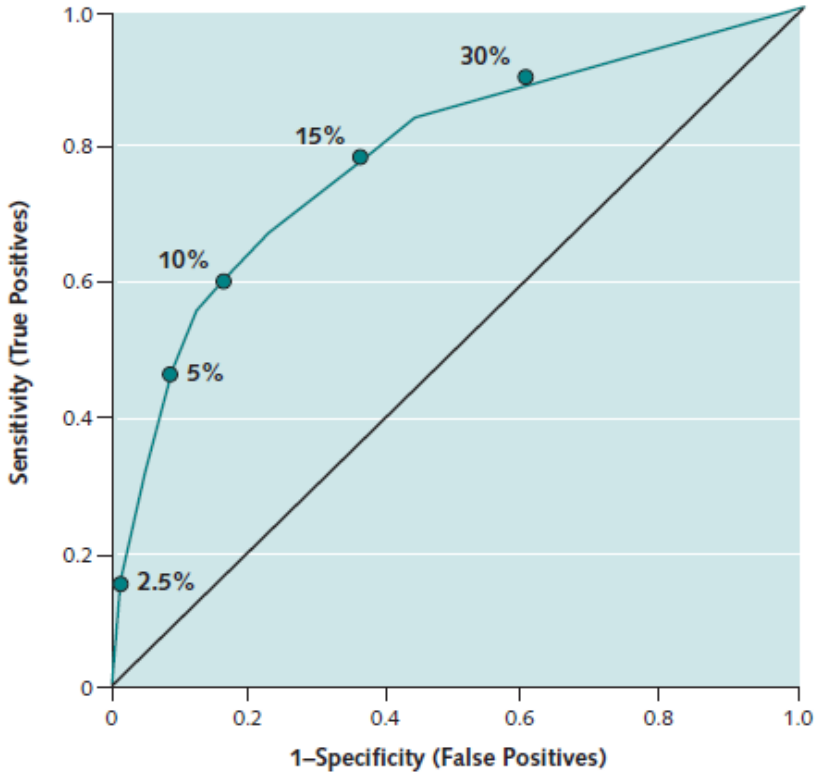
Figure 8. Example figure: a calibration plot with c-statistic and distribution of the predicted probabilities for individuals with and without the outcome (coronary artery disease).



Reproduced from reference 256, with permission from BMJ Publishing Group.

Model performance - discrimination

Figure 9. Example figure: a receiver-operating characteristic curve, with predicted risks labelled on the curve.



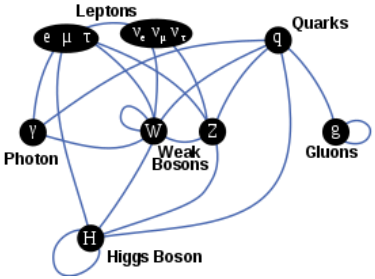
Receiver operating characteristic curve for risk of pneumonia . . . Sensitivity and specificity of several risk thresholds of the prediction model are plotted. Reproduced from reference 416, with permission from BMJ Publishing Group.

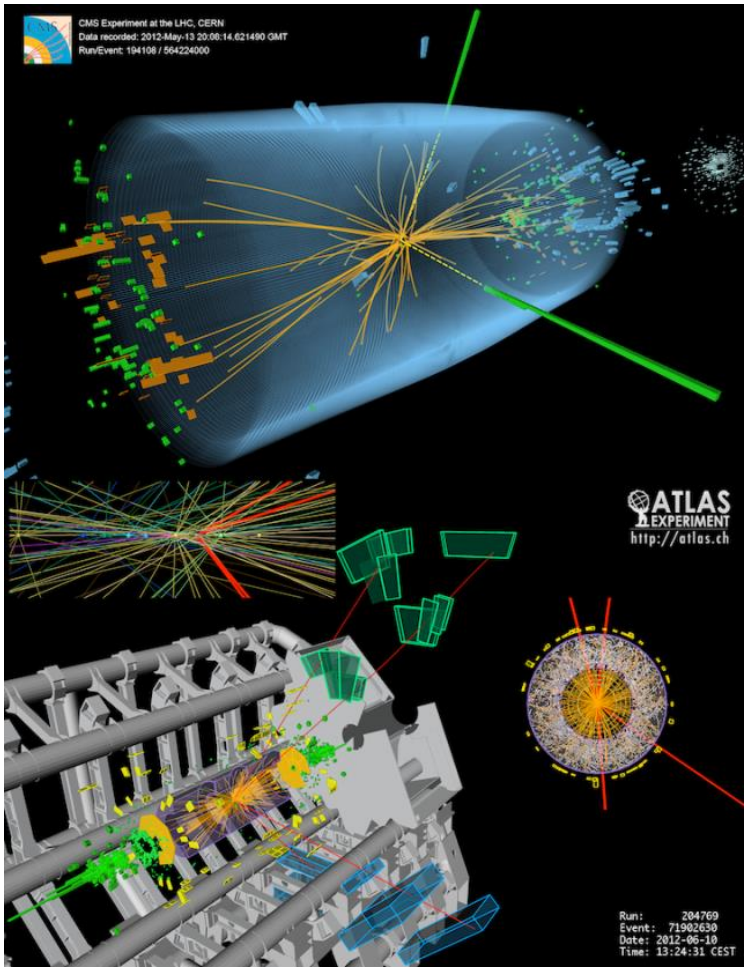
P-value

- The probability of getting the data observed or something more extreme, assuming the null hypothesis is true.
- (Note: P-values are r.v)
- This probability is *not* the probability that the null hypothesis is true.

Higgs boson (God particle)

- The Higgs boson is an elementary particle in the Standard Model of particle physics, produced by the quantum excitation of the Higgs field, one of the fields in particle physics theory. It is named after physicist Peter Higgs, who in 1964, along with five other scientists, proposed the mechanism which suggested the existence of such a particle.
- Its existence was confirmed in 2012 by the ATLAS and CMS collaboratio ns based on collisions in the LHCat CERN.





Wiki, https://home.cern/science/physics/higgs-boson

“the 5-sigma discovery criterion” = The probability that there's a Higgs is 99.9 percent?

- On 4 July 2012, CERN announced that they had discovered, at a significance of 5-sigma, a new particle with a mass consistent with that predicted for the Higgs boson.

- [one-sided test; 1- pnorm(5) = 1- 0.9999997 = 0.0000003]

- [The Independent, July 2012]
- “The new particle is very near to the 5-sigma level of significance - meaning that there is less than one in a million chance that their results are a statistical fluke”

- [https://news.nationalgeographic.com/news/2012/07/120704-god-particle-higgs-boson-new-cern-science/]
- Speaking to a packed audience Wednesday morning in Geneva, CERN director general Rolf Heuer confirmed that two separate teams working at the Large Hadron Collider (LHC) are more than 99 percent certain they've discovered the Higgs boson, aka the God particle—or at the least a brand-new particle exactly where they expected the Higgs to be.

- *the CERN teams*
- *“CMS observes an excess of events at a mass of approximately 125 GeV with a statistical significance of five standard deviations (5 sigma) above background expectations. The probability of the background alone fluctuating up by this amount or more is about one in three million.”*

- the ATLAS group
- *“A statistical combination of these channels and others puts the significance of the signal at 5 sigma, meaning that only one experiment in three million would see an apparent signal this strong in a universe without a Higgs.”*

Bayesian hypothesis testing

- Bayesian estimation (and confidence procedures)

- Prior probability: $P(H_0)$
- Posterior probability: $P(H_0|x)$

- the Bayes Factor : the posterior odds of H_0 relative to H_1
- [the Bayes Factor can be thought as the likelihood ratio, if the parameter spaces are simple values, the Bayes Factor is just the ratio of likelihoods under the two values.]

- $B_{01} = P(H_0|x)/P(H_1|x)$
- Reject H_0 If $B_{01} < 1$

A Comparison of Testing Methodologies
https://www2.stat.duke.edu/~berger/papers/testing-phystat.pdf
Could Fisher, Jeffreys and Neyman Have Agreed on Testing?

American Statistical Association (ASA) "Statement on Statistical Significance and P-Values"

- **1. P-values can indicate how incompatible the data are with a specified statistical model.**
  - A p-value provides one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data. The most common context is a model, constructed under a set of assumptions, together with a so-called "null hypothesis." Often the null hypothesis postulates the absence of an effect, such as no difference between two groups, or the absence of a relationship between a factor and an outcome. The smaller the p-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the p-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.

https://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN

- 2. **P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.**

  – Researchers often wish to turn a p-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The p-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.

- **3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.**
  - Practices that reduce data analysis or scientific inference to mechanical “bright-line” rules (such as “$p < 0.05$”) for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making. A conclusion does not immediately become “true” on one side of the divide and “false” on the other. Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis. Pragmatic considerations often require binary, “yes-no” decisions, but this does not mean that p-values alone can ensure that a decision is correct or incorrect. The widespread use of “statistical significance” (generally interpreted as “p 0.05”) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process

- **4. . Proper inference requires full reporting and transparency**
  - P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain p-values (typically those passing a significance threshold) renders the reported p-values essentially uninterpretable. Cherry picking promising findings, also known by such terms as data dredging, significance chasing, significance questing, selective inference, and “p-hacking,” leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided. One need not formally carry out multiple statistical tests for this problem to arise: Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis. Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all p-values computed. Valid scientific conclusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including p-values) were selected for reporting.

- **5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.**
  - Statistical significance is not equivalent to scientific, human, or economic significance. Smaller p-values do not necessarily imply the presence of larger or more important effects, and larger p-values do not imply a lack of importance or even lack of effect. Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive p-values if the sample size is small or measurements are imprecise. Similarly, identical estimated effects will have different p-values if the precision of the estimates differs.

- **6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.**
  - Researchers should recognize that a p-value without context or other evidence provides limited information. For example, a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis. Likewise, a relatively large p-value does not imply evidence in favor of the null hypothesis; many other hypotheses may be equally or more consistent with the observed data. For these reasons, data analysis should not end with the calculation of a p-value when other approaches are appropriate and feasible.