© 2025 Shizhe Chen. All rights reserved.

# Statistical Inference

> *"Statistics is the grammar of science."* – Karl Pearson

## Inference v.s. Prediction

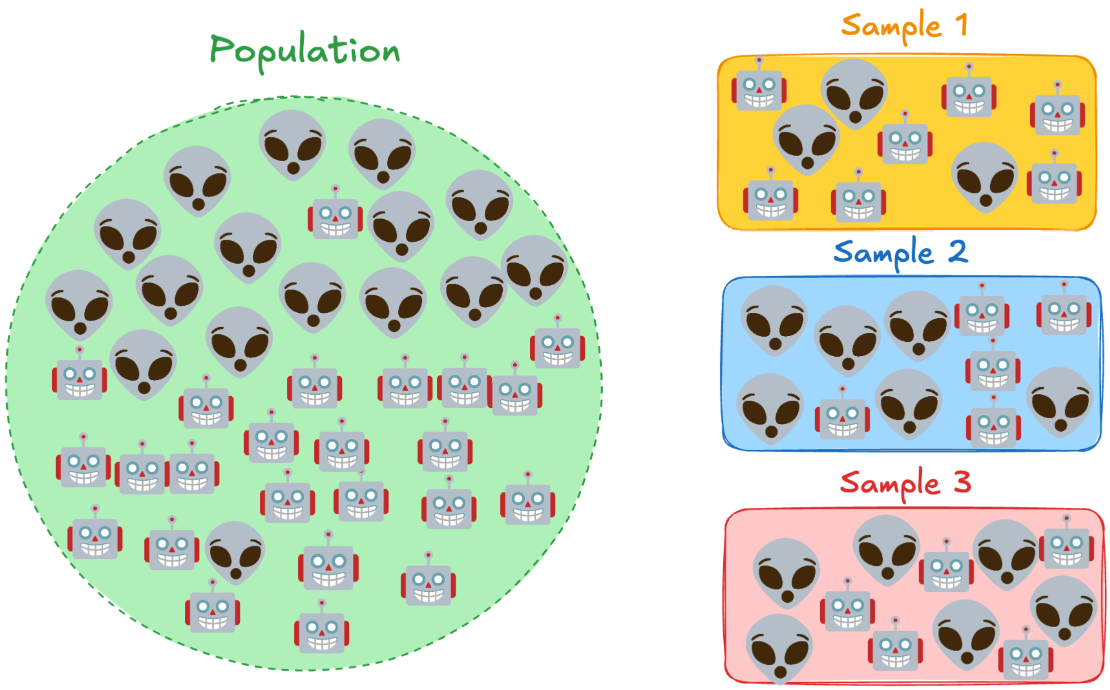
- Prediction: finding the best model that generalizes
- Inference: confirming a hypothetical (causal) relationship from data

## Bayesian v.s. Frequentist

- Bayesian: parameters are random but data are fixed
  Uncertainty directly available from the posterior distribution
- Frequentist: true parameters fixed but unknown
  Uncertainty as a result from having finite samples

## Population and Sample

What is the proportion of happy robots in the population, **if you only have access to one of the samples**?




Note: If you are able to acquire three sample sequentially, you can use the mark and recapture technique to estimate the size of the population.

## Sampling distributions

The distribution of a statistics under repeated sampling from the same population using the same sampling mechanism

- Acknowledging uncertainty of evidence from data
- Foundation of statistical reasoning
- Key to propert statistical inference

Example: Central Limit Theorem and Sampling Distribution

## Statistical Reasoning

Given that the observations are finite samples from a true distribution...

- Confidence Interval
  A set of estimators that guarantee chosen coverage probability.
- Hypothesis Testing
  Assuming the null hypothesis, is the dataset an extreme case?
- Multiple Testing
  If one tests infinite number of hypotheses, one of them might be positive by chance.
  xkcd#882 and explainxkcd

## The Magical Number $0.05$

Cowles and Davis (1982) wrote a note on *American Psychologist* regarding the origin of such practice:

- Fisher (1925): "It is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding **twice the standard deviation** are thus formally regarded as significant".
  Critical value = $2$ yields a two-sided p-value of $0.0456$.
- Prior to $2\ \mathrm{s.d.}$, $3\mathrm{PE}$ was used by William Gosset (Student 1908).
  $\mathrm{PE}$: probable error defined as semi interquartile range. ("Cumbrous, slipshod, and misleading phrase" Galton 1889.).
  $3\mathrm{PE} = 2.023\ \mathrm{s.d.}$ for standard nornal distribution.
- Prior to $3\mathrm{PE}$, Pearson (1900?) wrote '... p = .28 ("fairly represented" [p. 174]); p = .1 ("not very improbable ..." [p. 171]); p =.01 ("this very improbable result" [p. 172])'

Cowles and Davis: "Do people, scientists and nonscientists, generally feel that an event which occurs $5\%$ of the time or less is a rare event? Are they prepared to ascribe a cause other than mere chance to such infrequent events?"

So what do we think about $5\%$ in the era of big data and big models?

## The Magical Number $0.05$: Additional References

- The Magical Number Seven, Plus or Minus Two
- BASP banned NHSTP/p-values
- ASA statement on p-values
- Analysis of BASP after p-value ban
- The value of p
- Political Analysis banned p-value

## Other Common Tests

- Model-based Tests
- Rank-based Tests
- Simulation-based Tests

# Rank-based Tests

One way to avoid parametric assumptions is to consider the order/ranking instead of the actual numeric values.

Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ denote the ordered sample. The **rank** of an observation $x_i$ is

$$R_i \;=\; 1 + \#\{\, x_j : x_j < x_i \, j \neq i\}.$$

Bottom-line: For similar hypotheses, a parametric test is more efficient than a nonparametric test, while the nonparametric offers more robustness towards violation of assumptions.

## Model-based Tests

These tests are based on the comparison between two *nested* models, where the full model contain all parameters in the reduced model.

- F-test/ANOVA
- Likelihood ratio test
- ...

## Model-based Tests: F-test/ANOVA

Consider a linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}\big(\mathbf{0},\, \sigma^2\mathbf{I}\big)$$

Want to test $q$ linear constraints $H_0:\;\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$

Test statistic

$$F = \frac{\big(\mathrm{RSS}_\mathrm{r} - \mathrm{RSS}_\mathrm{f}\big)/q}{\mathrm{RSS}_\mathrm{f}/(n-p)}.$$

where $\mathrm{RSS}_\mathrm{r}$ is the residual sum of squares from reduced model and $\mathrm{RSS}_\mathrm{f}$ is the residual sum of squares from the full model.

Then under the null hypothesis $F \;\sim\; F_{q,\,n-p}$ *q* = 1 here.

## Likelihood Ratio test (LRT)

Using the logistic regression as an example

$$\Pr(Y = 1 \mid \mathbf{x}) \;=\; \mathrm{logit}^{-1}(\beta_0 + \beta_1 \mathrm{age} + \beta_2 \mathrm{smoke}).$$

To test $H_0 : \beta_2 = 0$ (“smoking not associated”):

$$\Lambda = -2\left[\log \mathcal{L}_{\mathrm{r}} - \log \mathcal{L}_{\mathrm{f}}\right],$$

where $\mathcal{L}_{\mathrm{r}}$ is the likelihood of the reduced model and $\mathcal{L}_{\mathrm{f}}$ is the likelihood of the full model.

Under $H_0$, $\Lambda \to \ \chi^2_{df=1}$ as sample size approaches infinity.

**Assumptions**

*Correctly specified likelihood* (independence, link function); sample size large enough.

> Note: For linear regression with normal errors, F-test is equivalent to LRT as sample size approaches infinity.

# Simulation-based Tests

Hypothesis testing hinges on the null distribution of a chosen test statistic.

When the closed-form null distribution is unavailable, we can approximate it.

- Simulation
  - Using Monte-Carlo methods to obtain the null distribution
  - Need to know the *true distribution*
  - Uncommon in practice but require no calculous
- Permutation
  - Shuffling data to break possible relationships
  - Work for specific null hypotheses
  - Limited applications but assumption-less
- Bootstrap
  - Approximate sampling distribution via resampling
  - Require large sample size
  - Computationally expansive but universal

© 2025 Shizhe Chen. All rights reserved.