DISTRIBUTIONS: THE SHAPE OF THINGS TO COME

DISTRIBUTIONS: WHEN YOU ARE SURE YOU HAVE THE WRONG ANSWER

- Imagine that we measure the heights, in inches (or any suitable unit) of 1000 UCSB students and this information only, what is the answer to the question

How tall will the next student you meet be?

- No matter how you choose to answer you are certain to be wrong.
  - The chances of the answer being exact are probabilistically 0
- Any answer you give summarizes the 1000 measurements in some way.

SOME DEFINITIONS

- Distribution: the structure or pattern of variation that a set of measurements of ***the same property*** exhibits.
- Statistic: a summarization of a distribution
- Estimate: A prediction of the ”true” value of a statistic computed from data
- Probability: a mathematical model describing a distribution or collection of distributions
- Independence: the notion that measurements or distributions have no probabilistic relationship

INFERENCE AND PREDICTION

- Inference: a statement, the truth of which is verified through mathematical analysis of data, about the properties (often expressed as a statistic) of a population.
- Prediction: a statement about an event that will occur in the future
- Who could ever confuse the two?

HOW TALL WILL THE NEXT STUDENT BE?

- If I give you 1000 measurements of UCSB student heights and ask you to predict the height of the next student you meet, what do you do?
- The standard procedure:
  - Treat the 1000 measurements as a random sample from the population of UCSB students
  - Use the average of the heights as a prediction of the height of the next student
- Why?

MEAN AND VARIANCE

- The mean is a statistic that summarizes a distribution
  - It is the "expected value" of a distribution computed as the sum (integral) of the value multiplied by its probability
- The variance measures the total squared distance between the mean and all values in a distribution
- The average (taken from a ***random sample*** from a population) is a good (unbiased) estimate of the population mean.
- Thus using the mean yields a good estimate of the squared distance.
- The distance is the "error" of the prediction. Thus the average is a prediction that minimizes the square error.

A WHOLE PILE OF ASSUMPTIONS

- **Random sample**: a subset of the members of the population chosen completely at random
- The mean is a summary of a probability distribution that is used to represent the distribution of values in the population
- The average is a good estimator of the mean
  - Some probability distributions do not have finite means
- The square error is the error measure to minimize

CONSIDER THE FOLLOWING RANDOM SAMPLE (SORTED)

- 1,2,3,4,5,6,7,8,9,10,101,102,103,104,105,106,107,108,109,110
- Predict that the next value drawn from the population will be the average:
- Avg: 55.65
- Is this a good prediction? It is far from ***any*** value in the list.

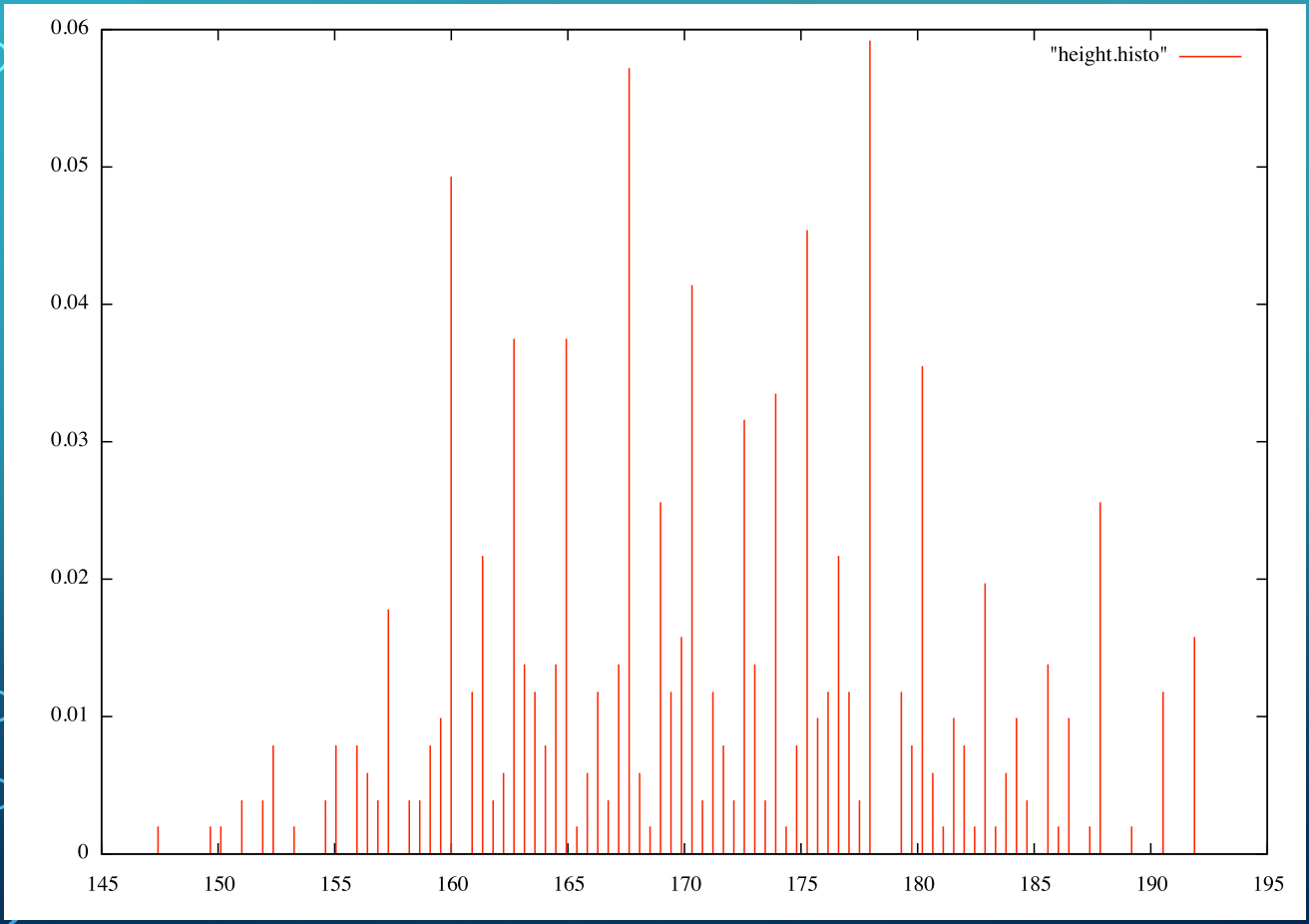
SUMMARIZING

- Inference and prediction are often confused
- The average is used to make an inference from a random sample about the mean of a probability distribution that you assume represents the distribution of the population.
- It is a good inference under very specific assumptions about the randomness of the sample
- There are cases when it is a really lousy prediction.

SOME REAL DATA

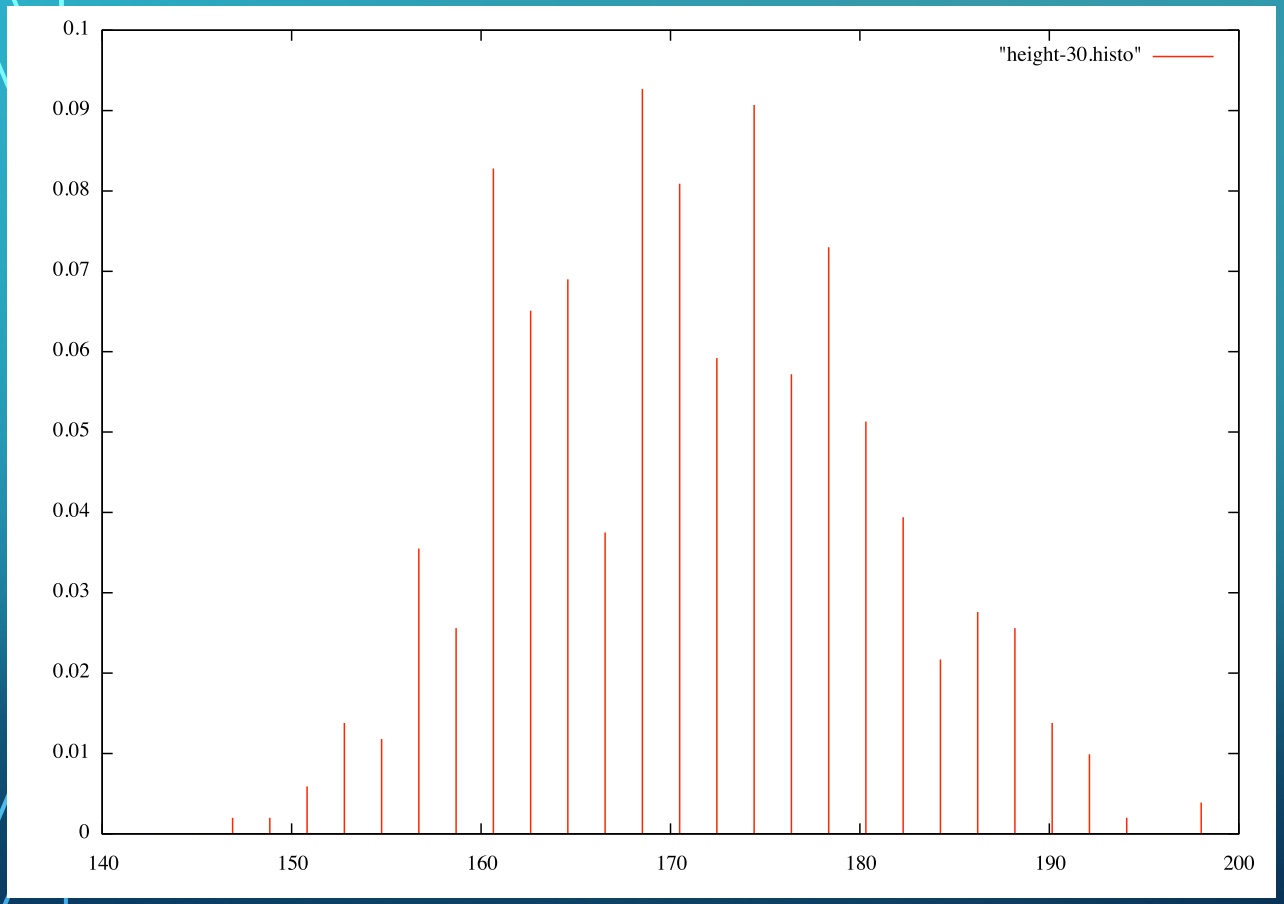
- https://www.statcrunch.com/5.0/shareddata.php
- Heights from 507 adults

HISTOGRAM OF THE HEIGHTS



- Is this data well modeled by a Normal (Gaussian) distribution?

HOW ABOUT NOW?

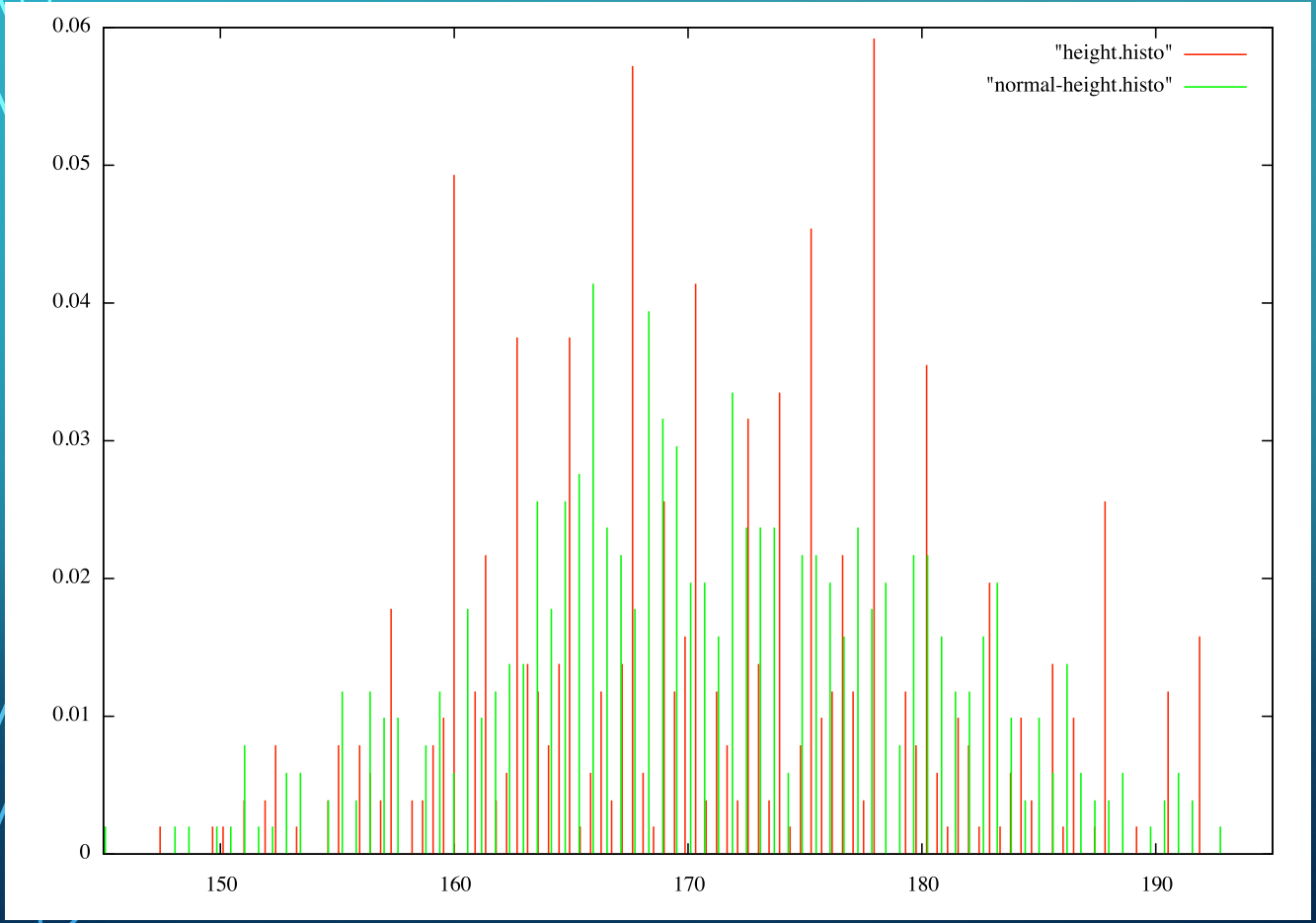


- 30 bins instead of 100

STATISTICS FROM THE DATA

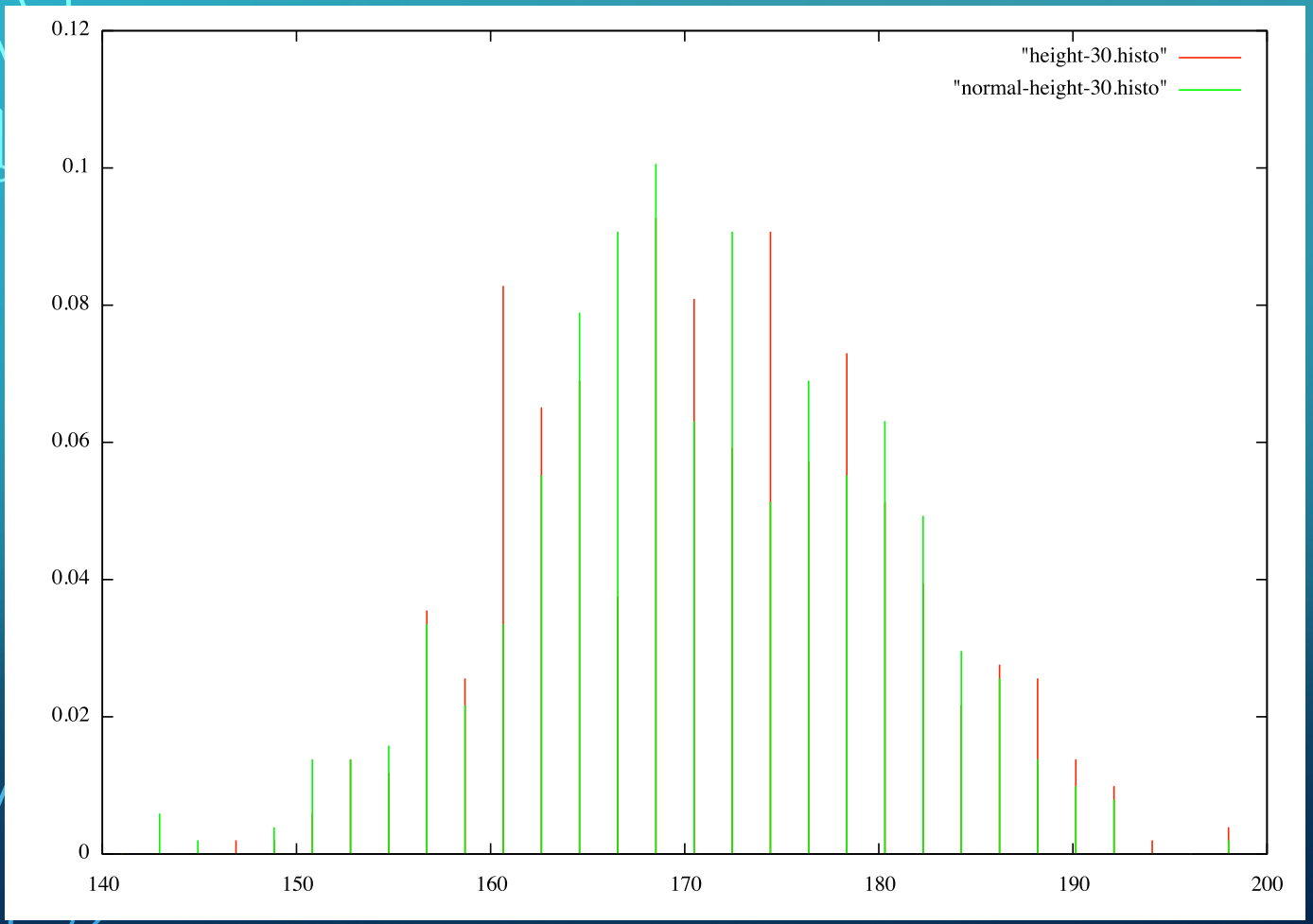
- Sample mean: 171.1 cm
- Sample standard deviation: 9.401 cm

HISTOGRAM OF DATA AND N(171.1,9.401)



- Are these two samples from the same distribution?

HOW ABOUT NOW?

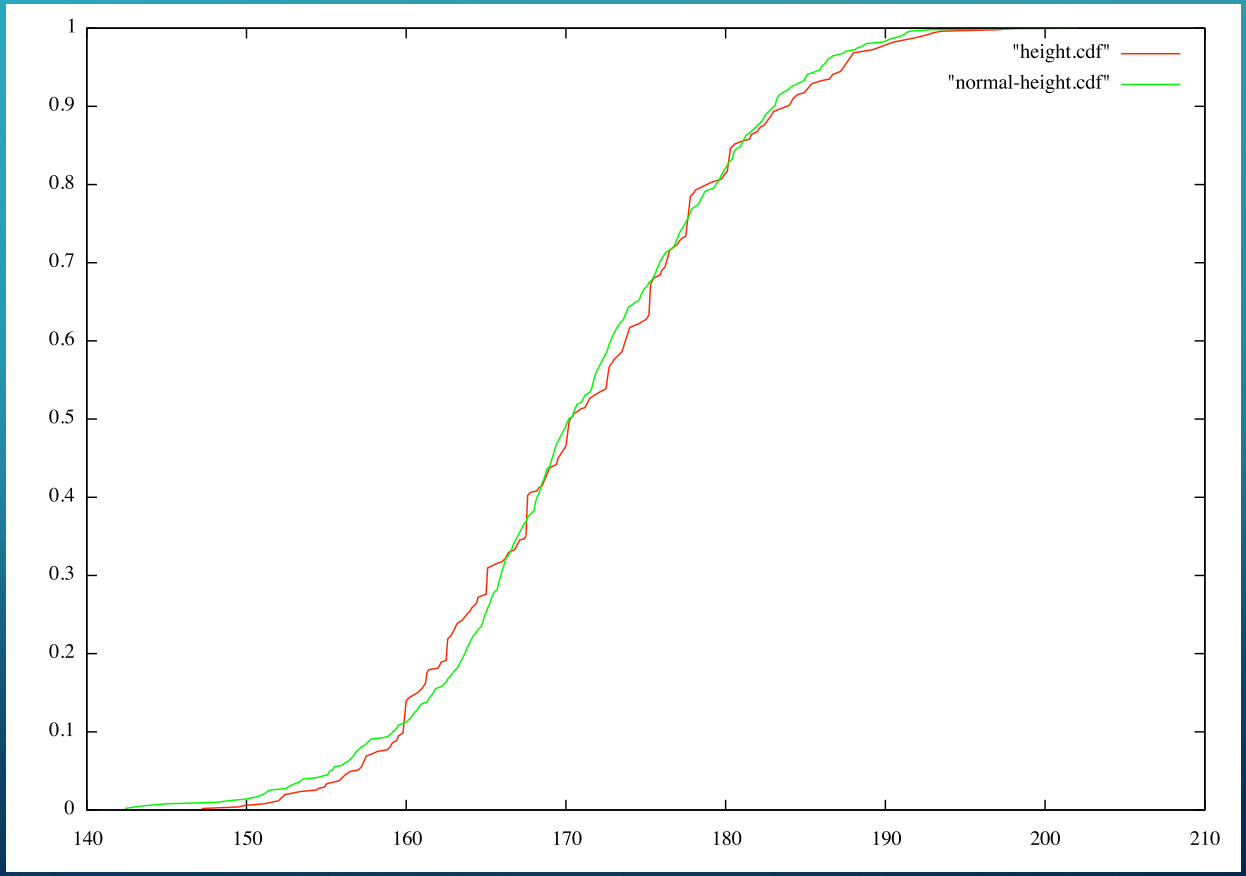


- Does your opinion of similar depend on bin size?

EMPIRICAL CUMULATIVE DISTRIBUTION FUNCTION

- X-axis shows the range of the data
- Y-axis shows the fraction of the data that is <= the corresponding x-axis value

EMPIRICAL CDF OF THE DATA AND N(171.1,9.401)



- Do these look similar?

VISUAL ANALYSIS

- Histograms are poor for visual analysis
  - Bin size matters
- Empirical CDF is usually the right tool for the job

THOUGHT PROBLEM

- Imagine you work for Amazon and they give you this data and tell you that you need to predict the “cut off” above which 5% of heights of the customers will fall (so they can charge extra shipping, let’s say, for large clothes). You get a bonus of $100K at the end of the year if the number if 5% to 3 digits. For every percentage point you are off, they dock your bonus $50K. What do you do to take home your bonus?