Mathematics 120: Lecture 7

Percentiles and Errors

Dan Sloughter

Furman University

September 17, 2018

Example

- From the 2004 Current Population Survey, the average family income was \$60,000 and the standard deviation was \$40,000.
- This means that the z-score for an income of 0 would be

$$\frac{0 - 60000}{40000} = -1.5.$$

- If the data were from a normal distribution, 6.7% of the families would have had a negative income.
- Reason: As we have seen, the distribution for incomes has a long right tail, and so is clearly not normal.
- Hence we cannot interpret the standard deviation in the same way as we do with data from a normal distribution.

Percentiles

- The *pth sample percentile* of a set of data is the value a for which p percent of the data is less than or equal to a.
- Note:
  - The median is the 50th percentile.
  - The 25th and 75th percentiles are also called the *first* and *third quartiles*, respectively.
  - That is, the first quartile is the median of the first half of the data; the third quartile is the median of the second half of the data.
- We call the distance between the first and third quartiles the *interquartile range*.
- That is, the interquartile range is the length of the interval containing the central 50% of the data.

Example

- The bat-to-insect data values in order: $23, 34, 40, 42, 45, 52, 56, 62, 68, 83$.
- Recall: The median is 48.5.
- The first quartile is 40 and the third quartile is 62.
- The interquartile range is 22.

Example

- Note: Saying that the area under the standard normal curve from -2 to 2 is about 0.95 is equivalent, by symmetry, to saying that the 97.5 percentile is about 2.
- More precisely, the 97.5 percentile of the standard normal is 1.96.
- Using the table, we find that the 80th percentile is about 0.85

Example

- Suppose certain standardized test scores follow a normal curve with mean 500 and standard deviation 100.
- Q: What is the 90th percentile?
- The 90th percentile of a standard normal is about 1.30.
- So the score we are looking for is $1.30 \times 100 = 130$ points above the mean.
- Hence the 90th percentile is $500 + 130 = 630$ points.

Errors in measurements

- In the real world, repeated measurements of the same quantity do not yield the same results.
- Example
  - In astronomy, repeated astronomical observations of a planet or other object will yield slightly differing results.
  - Astronomers developed mathematical tools for dealing with the errors in measurement.
  - In particular, Gauss used the normal curve in order to understand the behavior of the errors.
- Note: Although we might have some idea of what leads to the measurement errors, the errors are themselves unknowable.
- That is, the only way we can account for the errors is through chance models, that is, mathematical models based on probability.
- Example: Standard weights (see Section 6.2 in the text).

Chance model

- For any measurement, we assume the quantity being measured has some true exact value, which we might denote by μ.
- The error of an individual measurement, say ϵ, is then the difference between the measured valued and the exact value.
- That is, if x is an observed individual measurement, we have

$$x = \mu + \epsilon.$$

- That is, in words,

$$\text{individual measurement} = \text{exact value} + \text{chance error}.$$

- Note
  - x is what we know; that is, it is what we actually observe.
  - μ is fixed, but unknown and unknowable (at least not knowable exactly).
  - ϵ is random, and varies from observation to observation.
  - ϵ is also not known (if it were, we would also know μ).

Chance model (cont'd)

- For most measurement data (for example, measuring astronomical distances or weighing an object), the error in the measure will have a normal distribution with mean 0 and some standard deviation σ.
- Equivalently, the measurements themselves have a normal distribution with mean μ and standard deviation σ.
- In this case, σ provides a measure of the magnitude we might expect in the chance error.
- For example, for the standard weight example in the text, repeated measurements of the standard weight provide an approximation of σ (using the standard deviation SD of the measurements), which then provides an estimate of the magnitude of the chance error when comparing the standard weight with another weight.

Outliers

- An *outlier* in a set of data is an observation which differs significantly from the rest of the data.
- Question about any outlier: Is it real, or is it the result of some non-random error in the data collection.
- An outlier could be the result of, for example,
  - an observation from a subject which is not part of the population under study,
  - a change in the measurement mechanism, or
  - an error in data entry.
- Outliers are sometimes thrown out from data sets, but one must do so only if there is sufficient evidence to show that it is not truly representative of the population.

Bias

- *Bias* is a systematic, non-random error introduced into the measurements.
- For example, a standard weight may be known to be off from the internationally accepted standard weight.
- If bias is present, our chance model changes to

$$\text{individual measurement} = \text{exact value} + \text{bias} + \text{chance error}.$$

- Note:
  - If known, one may adjust the data for the bias.
  - However, one cannot discover a bias from the data alone.