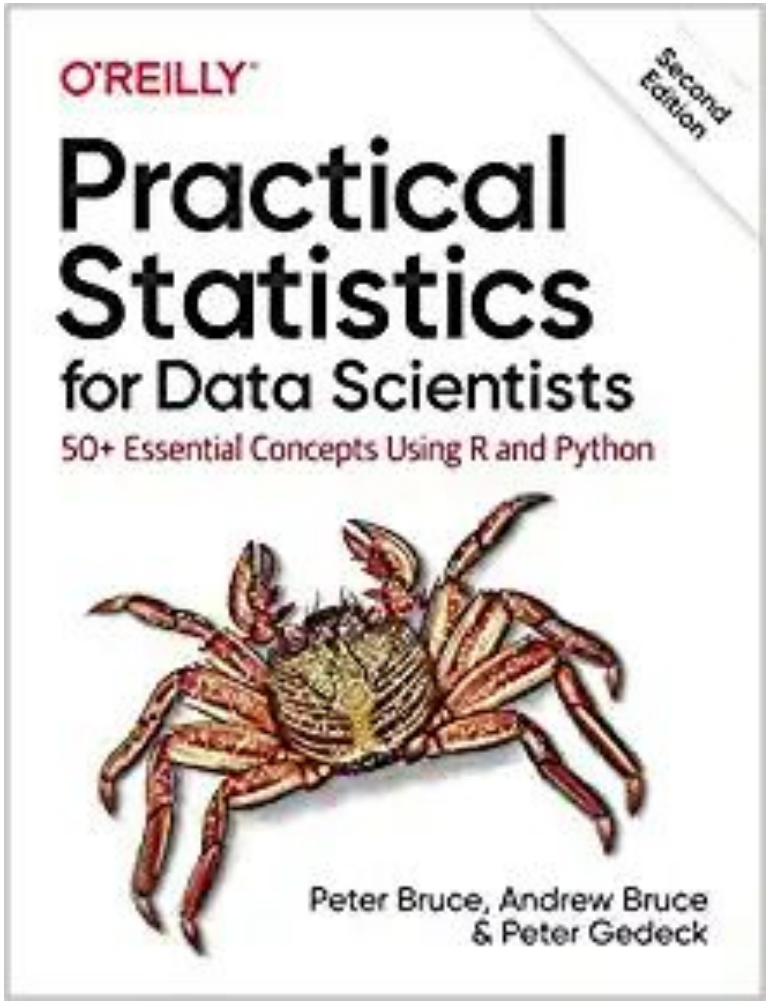


Chapter 2: Data and Sampling Distributions

Chapter content

Random Sampling and Sample Bias

- *Sample* subset from a larger dataset, the *population*.
- *Random sampling* each sample of the population has an equal chance to be sampled. The sampling can be done with or without replacement.
- *Stratified sampling* Dividing the population into strata and randomly sampling from each strata.
- *Bias*: Systematic error.
- *Sample bias*: A sample that misrepresents the population.

Selection Bias

- *Selection bias*: Bias resulting from the way in which observations are selected.
- *Data snooping*: Extensive hunting through data in search of something interesting.
- *Vast search effect*: Bias or non-reproducibility resulting from repeated data modeling, or modeling data with large numbers of predictor variables.

  If you repeatedly run different models and ask different questions with a large data set, you are bound to find something interesting. But is the result you found truly something interesting, or is it the chance outlier?

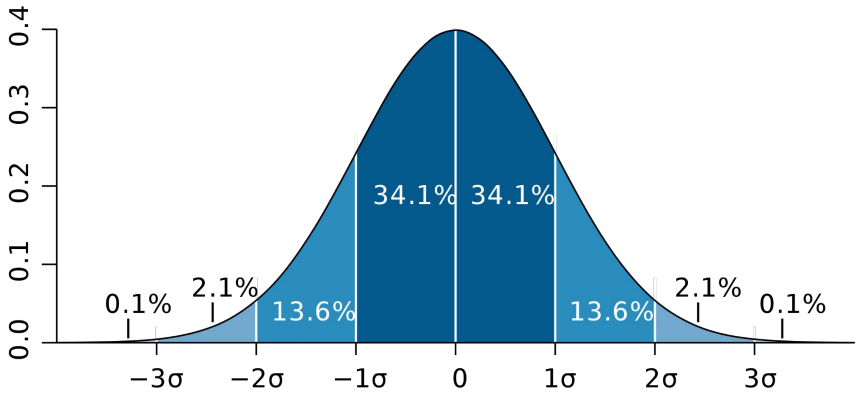
Sampling Distribution of a Statistic

To study a population, we draw a sample to measure its statistics: mean, medians, quartiles. The metrics differ from the whole population and samples to samples.

Sampling Distribution: distribution of some sample statistic over many samples drawn from the same population.

The Sampling distribution of a statistics for example the mean will often differ from the distribution of the population.

Normal Distribution



Source: Wikipedia

The Normal Distribution is a key tool in statistics due to the **Central Limit theorem.**
Although many population distribution are not normally distributed the sampling distribution of the mean is.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

μ: population mean
σ: Population variance

StatQuest: The Normal Distribution, Clearly Explained

Sampling Distribution of a Statistic

Central Limit Theorem: when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a bell curve) even if the original variables themselves are not normally distributed.

Therefore the distribution of the mean from many samples will be normally distributed even if the whole population is not normally distributed. Therefore if your samples are large enough (> 30) the normal distribution can be used to study the mean of a samples.

In Data Science this can be relevant for A/B testing. The Bootstrap technique is often more relevant to data science problems.

StatQuest: The Central Limit Theorem

https://en.wikipedia.org/wiki/Central_limit_theorem

Sampling Distribution of a Statistic

Standard error: a metric that sums up the variability in the sampling distribution for a statistic. The standard error should not be confused with the standard deviation.

The Standard error can be estimated using the bootstrap resampling technique.

$$SE = \frac{s}{\sqrt{n}}$$ with s: standard deviation.

The Bootstrap

- Estimate the sampling distribution of a statistics by picking samples with replacement from the dataset.
- Computationally expensive but powerful tools which do not rely on a any assumption on the population.
- It allows to estimate how a given statistics would evolve if we collected more samples from the population.

See StatQuest videos:

- https://www.youtube.com/watch?v=Xz0x-8-cgaQ
- https://www.youtube.com/watch?v=N4ZQQqyIf6k

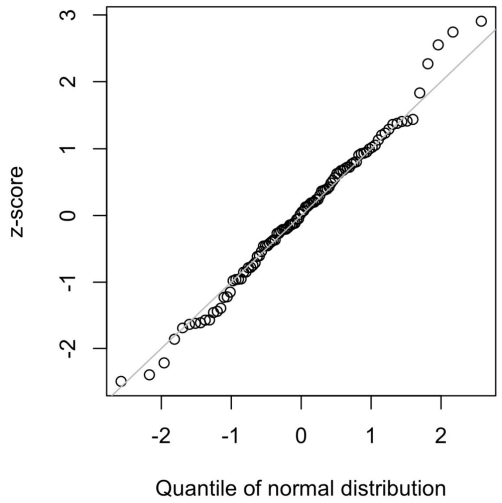
Confidence Intervals

1. “The confidence interval (CI) is a range of values that's likely to include a population value with a certain degree of confidence.” (see here)
2. The confidence intervals can be estimated using the Bootstrap method.
3. Estimation methods:
    1. Resample your data with replacement
    2. Record the statistic of interest (for ex the mean)
    3. Repeat 1 and 2 R time
    4. For a 95% confidence interval trim the results using the formula: [(100-95) / 2]. The trimmed sampling distribution is the bootstrap confidence interval estimation

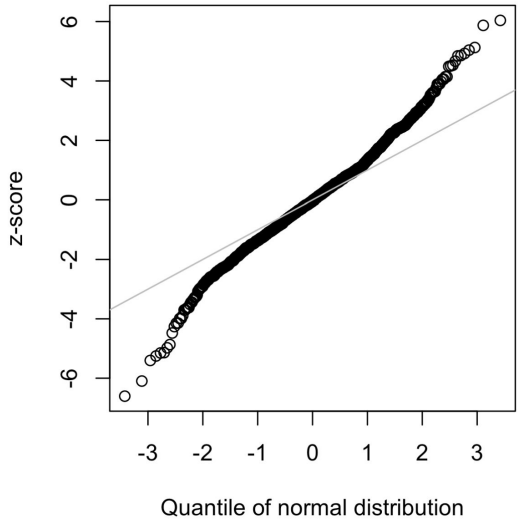
Long-Tailed Distributions

- Most data are not normally distributed. Extreme values with low frequency are present in many distributions. They form the *tails* of the dataset.
- Example stock values, house prices.

Normally distributed dataset



Long tailed dataset



Student's t-Distribution

- Published in 1908 in Bio- metrika by W. S. Gosset under the alias "Student."
- Approximate sampling distributions of a sample statistics when n are small.
- If n is large the t-Distribution is the classical Normal distribution.
- The t-Distribution is used to compare the mean of different samples.

Binomial Distribution

- The binomial distribution is used to model situation with 2 outcomes. Flipping a coin, Buy/don’t Buy, Die/Survive.
- If n is large and no outcome probability is close to 0 or 1 then the distribution can be modelled by the Normal Distribution.

Chi-Square Distribution

- The Chi-Square Distribution is used to compare the proportions of categorical variables between populations.

F-Distribution

- Used to compare the effect of multiples treatments across different groups.
- The F-statistics is used in the analysis of variance ANOVA test.

Poisson and Related Distributions

- Poisson distribution: used to model events occurrence across time or space.
- Exponential Distribution: used to model interval of time between events.
- Weibull Distribution: used to model event occurrence when the event probability change with time. For example occurrence of failure in a car engine.

This was a summary created by the members of the #66DaysOfData Study Group William Guesdon and Rea Kalampaliki.

#66DAYSOFDATA