PRACTICAL STATISTICS FOR DATA SCIENCE

REVIEW

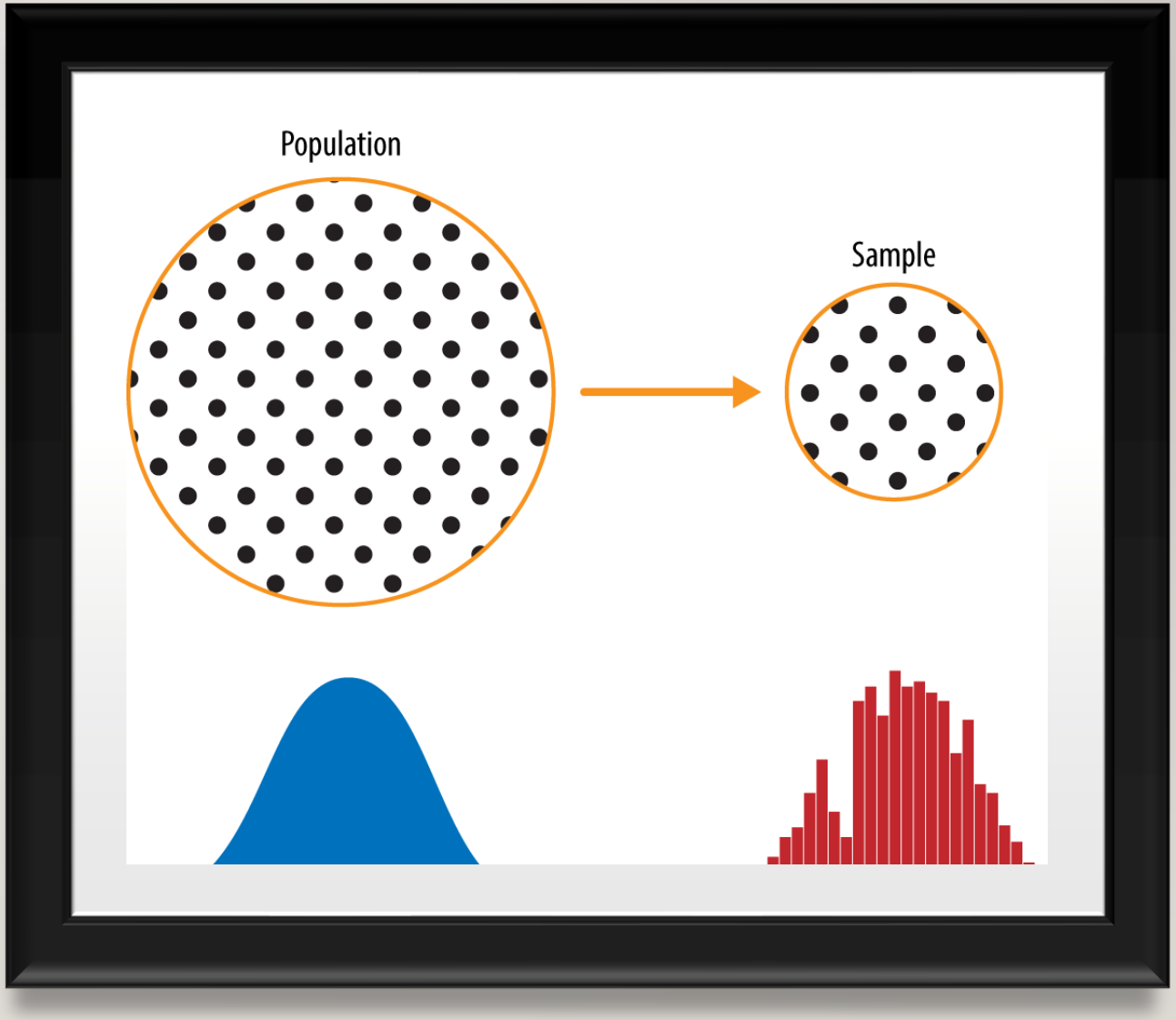
EXPLORATORY DATA ANALYSIS

CHAPTER 2

Data sampling and Distributions

KEY CONCEPTS

- **Population:** The larger dataset or idea of a dataset.
- **Sample**: A subset from a larger dataset.
- **N(n):** The size of the population sample.

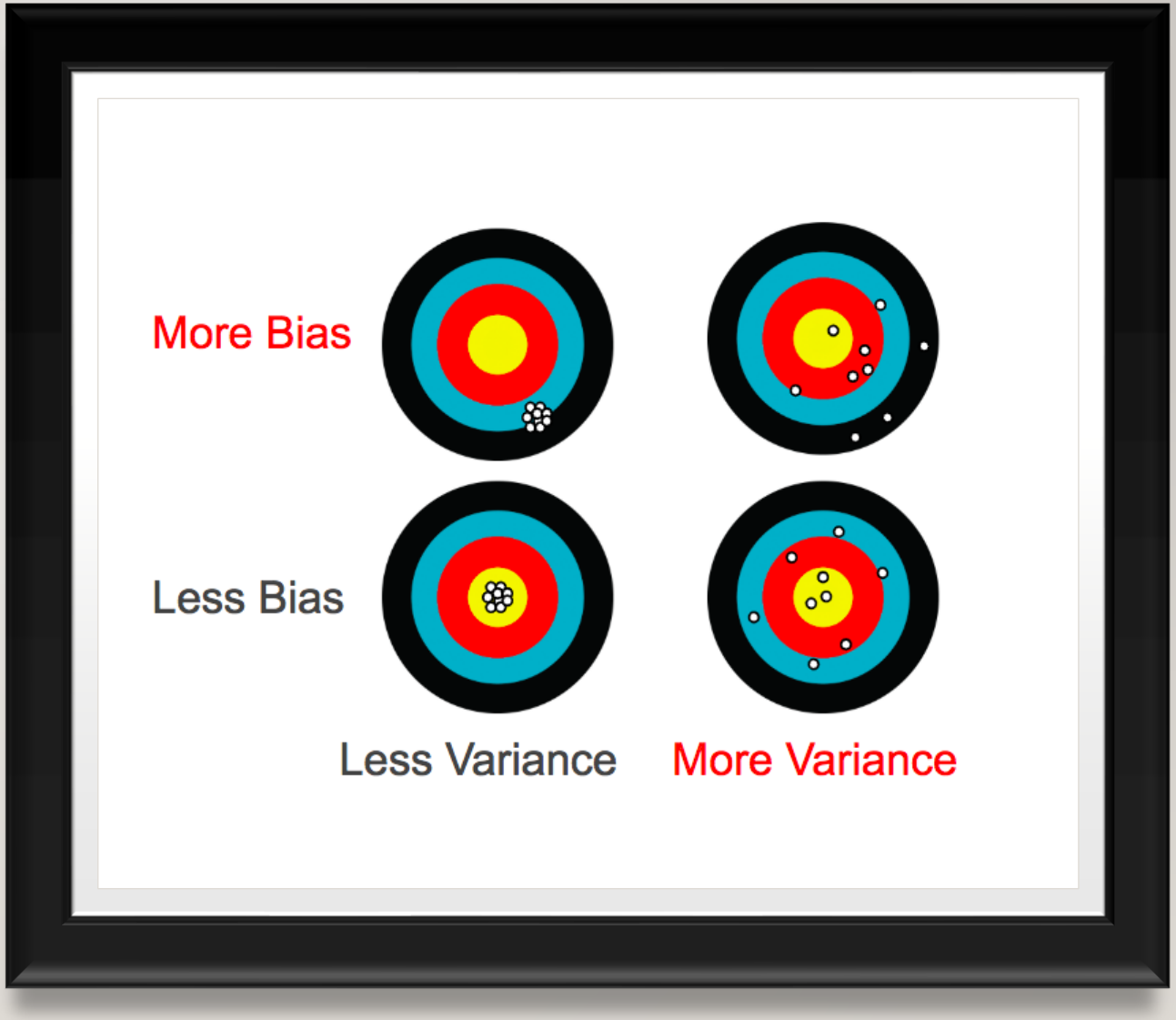


RANDOM SAMPLING AND BIAS

- **Random sampling**: Drawing elements into a sample at random. Each available member of the population being sampled has an equal chance of being chosen for the sample at each draw.
  - With Replacement: observations are put back in the population after each draw for posible future reselection.
  - Without replacement: observations, once selected, are unavailable for future draws.
- **Sample Bias**: A sample that misrepresents the population. (Poll Example)

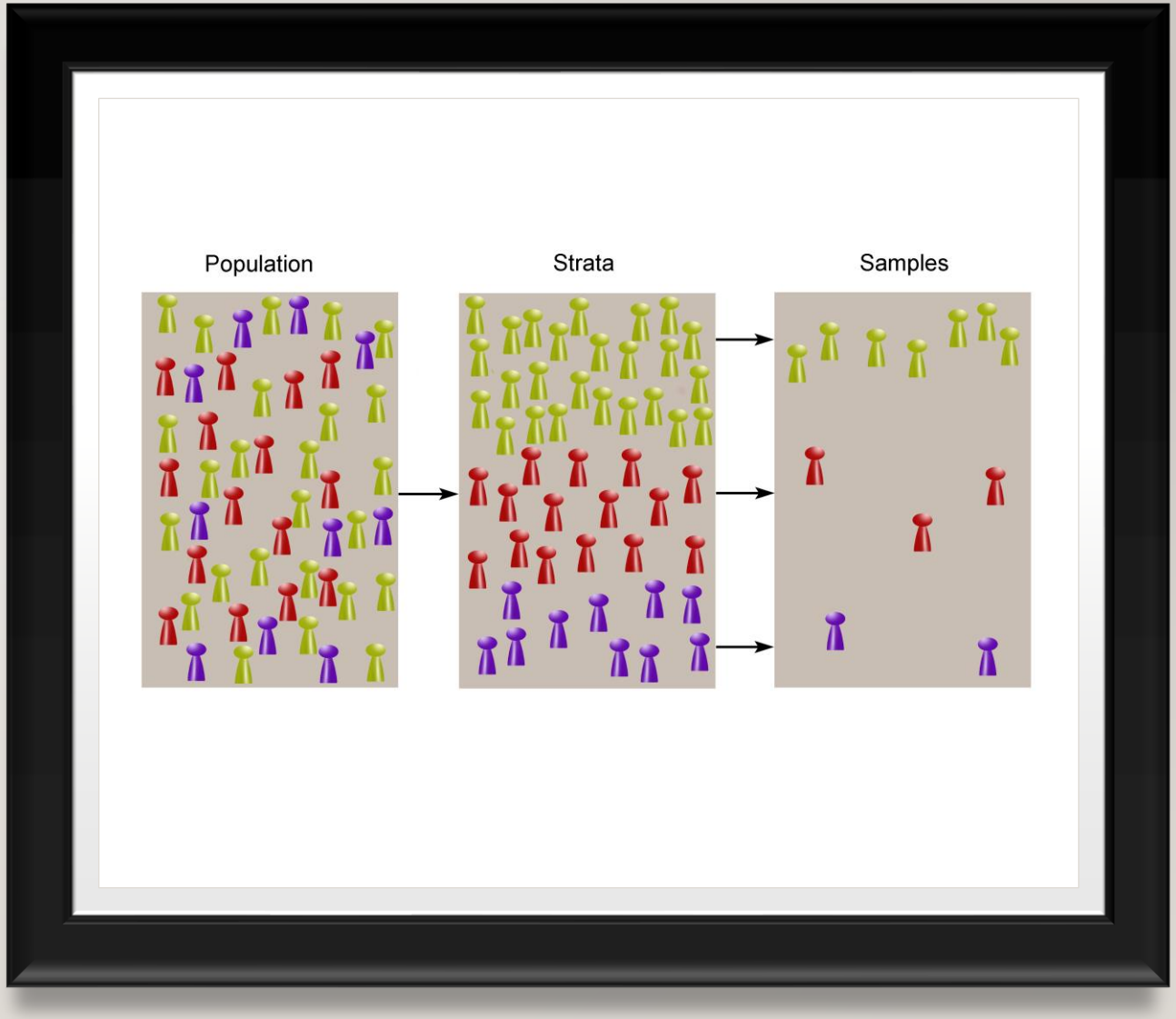
BIAS

- Statistical bias refers to measurement or sampling errors that are systematic and produced by the measurement or sampling process.



RANDOM SELECTION

- In **stratified sampling**, the population is divided up into strata, and random samples are taken from each stratum.

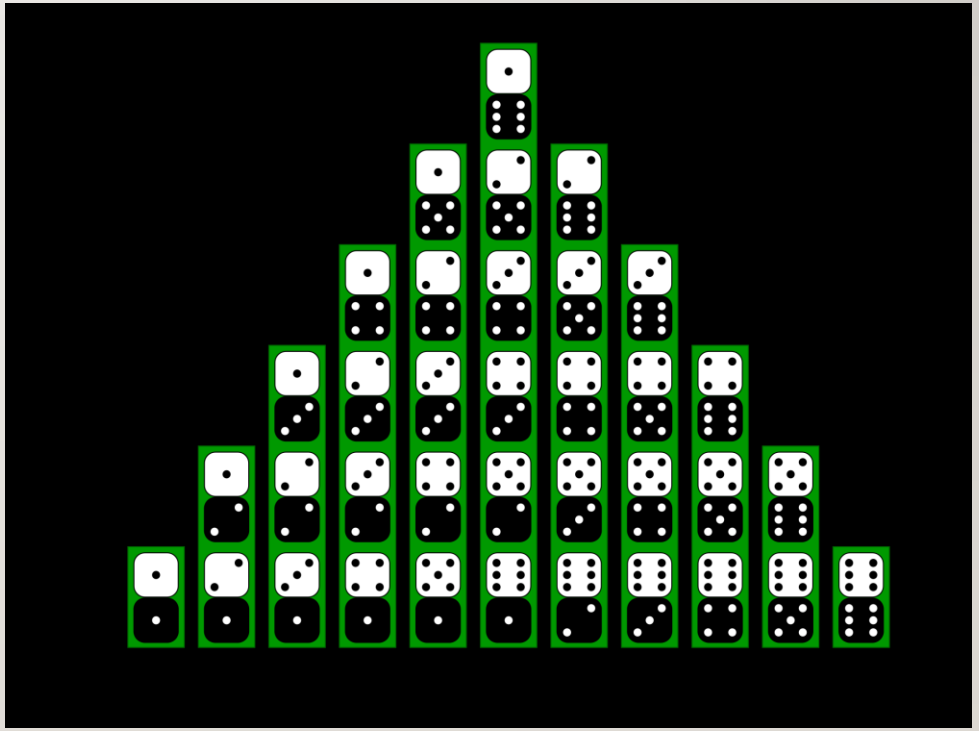


SAMPLING KEY IDEAS

- Even in the era of big data, **random sampling** remains an important arrow in the data scientist's quiver.
- **Bias** occurs when measurements or observations are systematically in error because they are not representative of the full population.
- Data quality is often more important that data quantity, and random sampling can reduce bias and facilitate quality improvement that would be prohibitively expensive.

SELECTION BIAS

- **Data snooping**: Extensive hunting through data in search of something interesting. "If you torture the data long enough , sooner or later it will confess."
- **Vast search effect**: Bias or non-reproducibility resulting from repeated data modeling, or modeling data with large numbers of predictor variables.
- **Regression to the mean**: refers to a phenomenon involving successive measurements on a given variable: extreme observations tend to be followed by more central ones.



SAMPLING DISTRIBUTION OF A STATISTIC

- The **sampling distribution** of a statistic refers to the distribution of some sample statistic, over many samples drawn from the same population.
- Typically, a sample is drawn with the goal of measuring something or modeling something. We are interested on ***sampling variability***.
- The distribution of a sample statistic such as the mean is likely to be more regular and bell-shaped than the distribution of data itself.

LAW OF LARGE NUMBERS

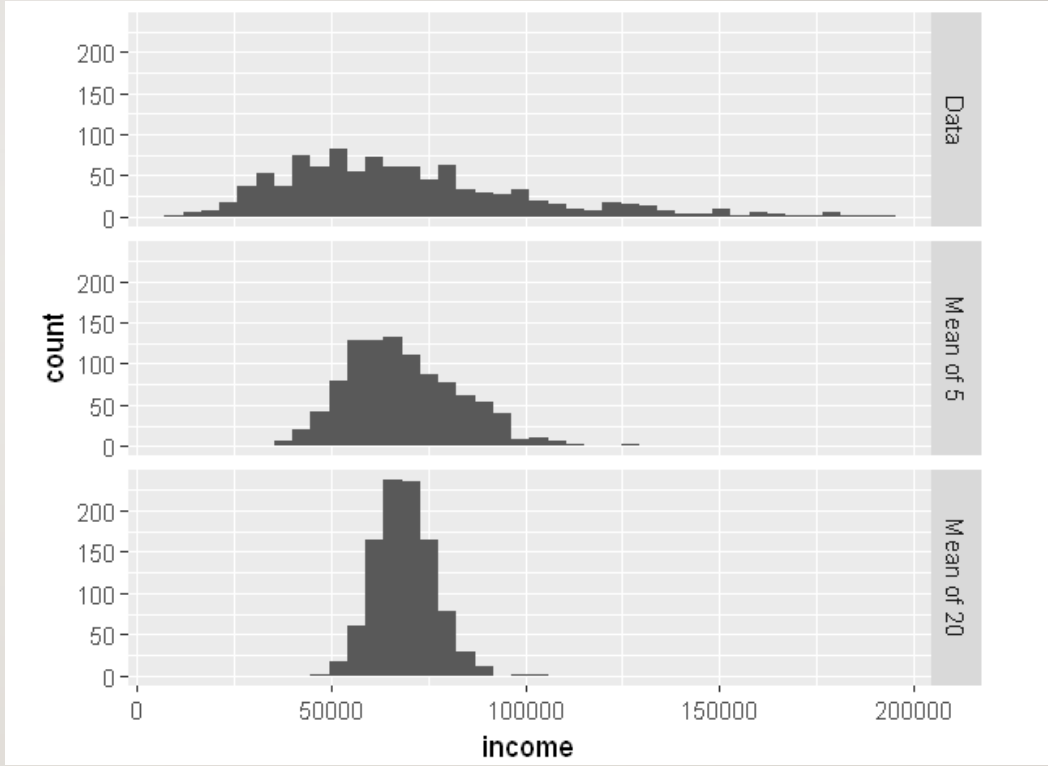
- The average of the results obtained from a large number of trials should be close to the expected value and will tend to become closer to the expected value as more trials are performed.

$$\overline{X}_n \xrightarrow{P} \mu \qquad \text{when } n \to \infty.$$

That is, for any positive number ε,

$$\lim_{n\to\infty} \Pr\left(|\overline{X}_n - \mu| > \varepsilon \right) = 0.$$

- The distribution of a sample statistic such as the mean is likely to be more regular and bell-shaped than the distribution of data itself.



CENTRAL LIMIT THEOREM

- The sampling distribution of the mean approaches a normal distribution, as the sample size increases.

- Proof*

$$z = \sqrt{n}\frac{\overline{x} - \mu}{\sigma} \longrightarrow_d \mathcal{N}(0,1) \; as \; n \to \infty.$$

STANDARD ERROR

- Single metric that sums up the variability in the sampling distribution for a statistic. The standard error can be estimated using a statistic based on the standard deviation s of the sample values, and the sample size n:

$$SE = \frac{s}{\sqrt{n}}$$

NOTE: standard deviation measures the variability of individual data points and standard error measures the variability of a sample metric

THE BOOTSTRAP

- An effective way to estimate the sampling distribution of a statistic, is to draw additional samples, with replacement, from the sample itself and recalculate the statistic.
  - Is a powerful tool for assessing the variability of a sample statistic.
  - Can be applied in a wide variety of circumstances, without study of mathematical approximations to sampling distributions.
  - Estimate sampling distributions for statistics where **no mathematical approximation has been developed.**
  - When applied to predictive models, aggregating multiple bootstrap sample **predictions outperforms the use of a single model.**

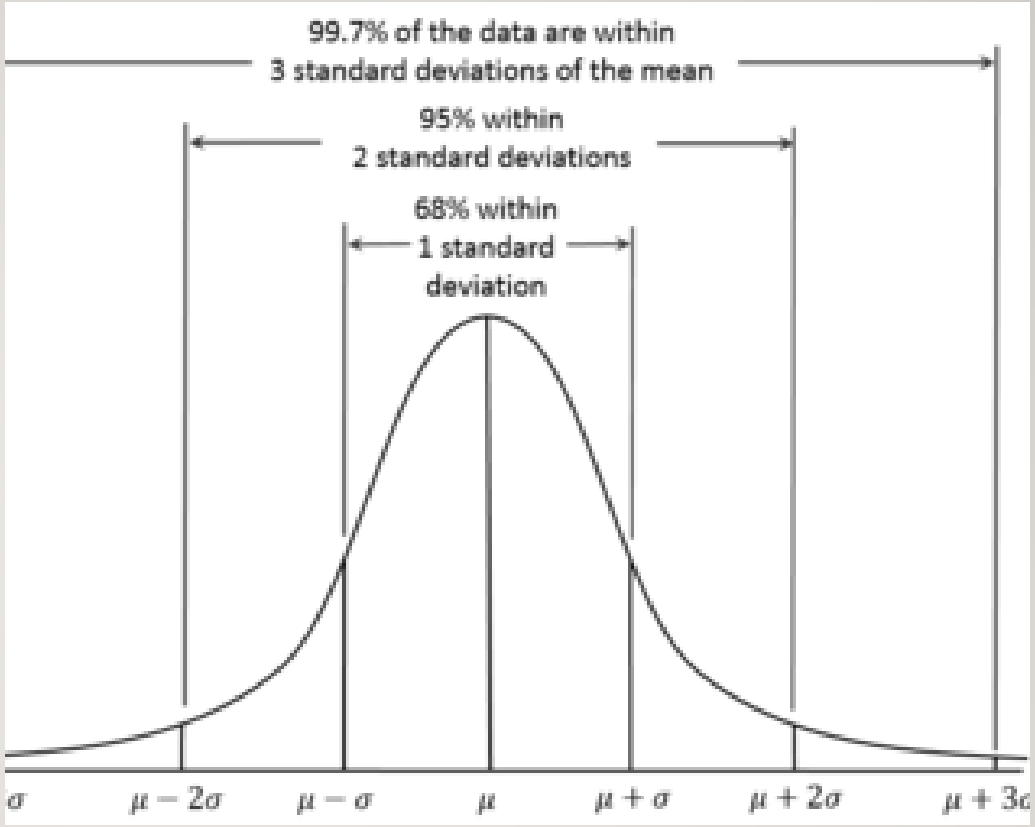
BOOTSTRAP OF THE MEAN ALGORITHM

1. Draw a sample value, record, replace it.
2. Repeat n times.
3. Record the mean of the n resampled values.
4. Repeat 1-3 B times
5. Use the B results to :
    a) Calculate their SD.
    b) Produce a histogram or boxplot.
    c) Find confidence interval.

CONFIDENCE INTERVALS

- Confidence intervals is an alternative to point estimation. It is a good way to deal with uncertainty. Confidence interval are the typical way to present estimates as an interval range.
- The more data you have, the less variable a sample estimate will be.
- The lower the level of confidence you can tolerate, the narrower the confidence interval will be.
- Bootstrap is an effective way to construct confidence intervals.

NORMAL DISTRIBUTION



- Bell-shaped distribution, Gaussian distribution.

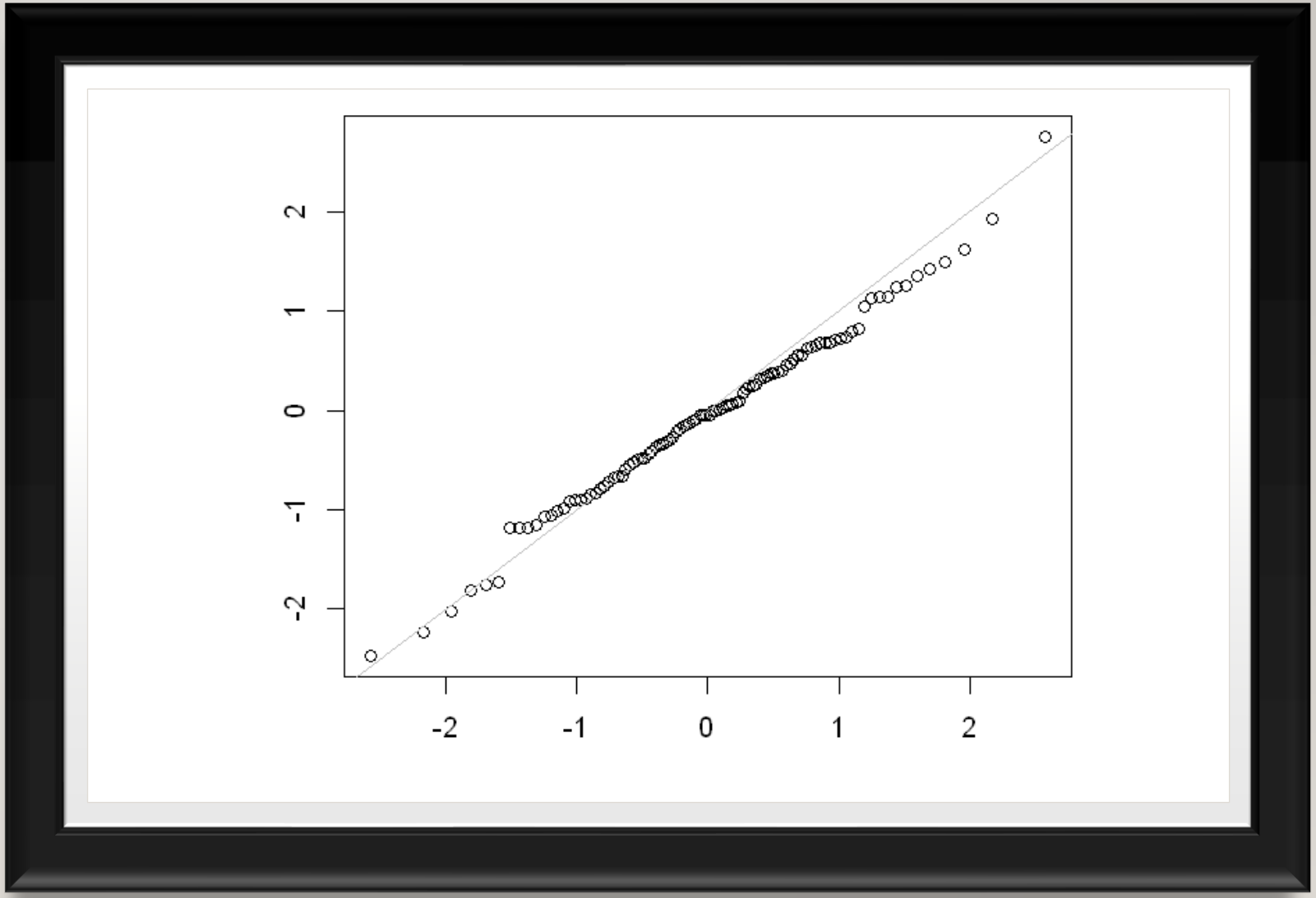
$$\phi_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

STANDARD NORMAL

- Let $\theta \sim N(\mu, \sigma^2)$ normal distributed.
- $z = \frac{\theta - \mu}{\sigma} \sim N(0,1)$ This transformation is commonly called standardization or z-scores.

- Note: Converting data to z-scores does not make the data normally distributed. It just puts the data on the same scale as the standard normal distribution.

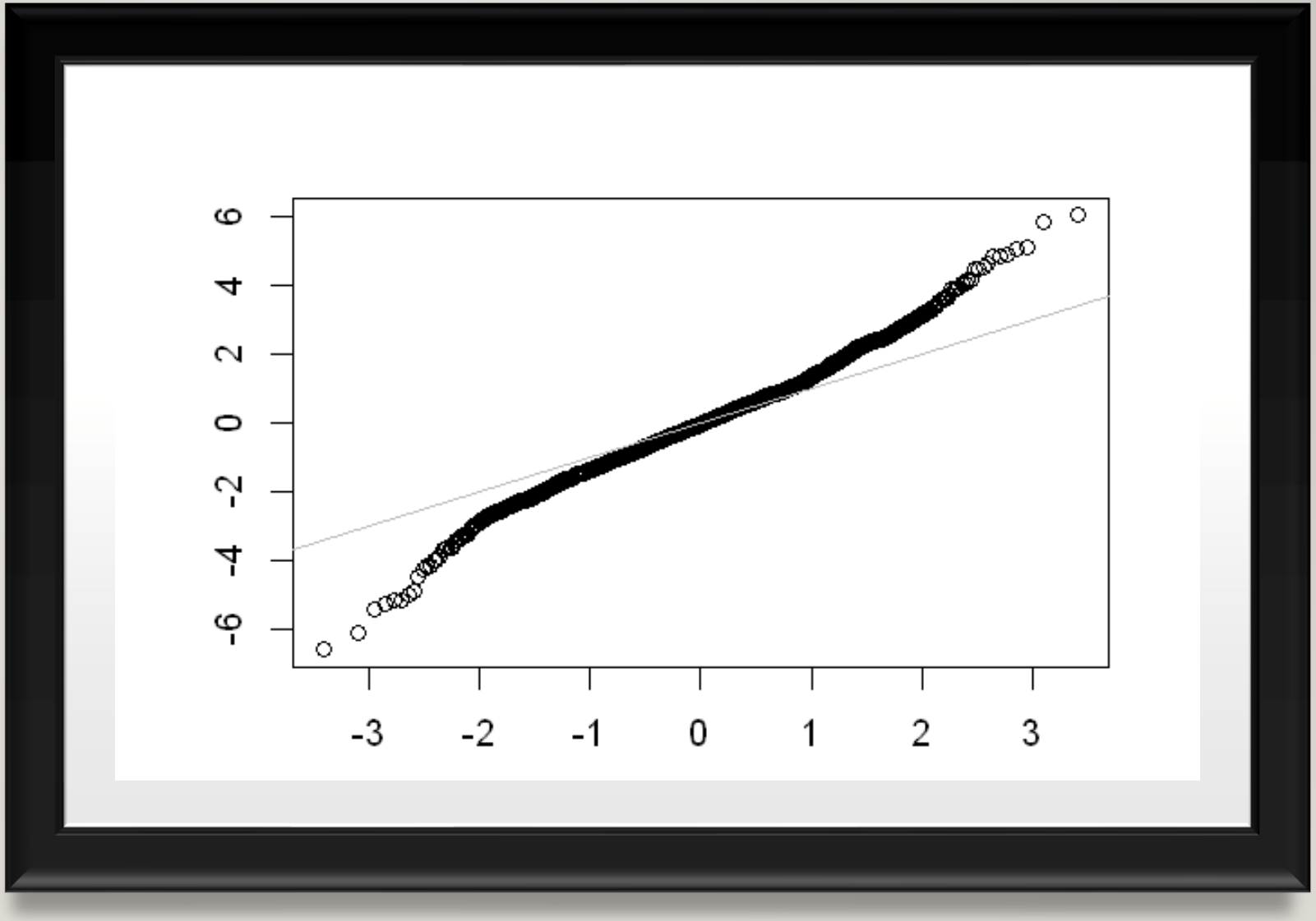
QQ-PLOTS



LONG-TAILED DISTRIBUTIONS

- Data is generally not normally distributed!!!
- **Tail**: The long narrow portion of a frequency distribution, where relatively extreme values occur at low frequency.
- **Skew**: Where one tail of a distribution is longer than the other. (Asymmetry)

NETFLIX STOCKS QQ-PLOT



STUDENT'S T-DISTRIBUTION

- The t-distribution is a normally shaped distribution, but a bit thicker and longer on the tails. Often called Student's t.

$$f(t) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}(1+t^2/\nu)^{-(\nu+1)/2}$$

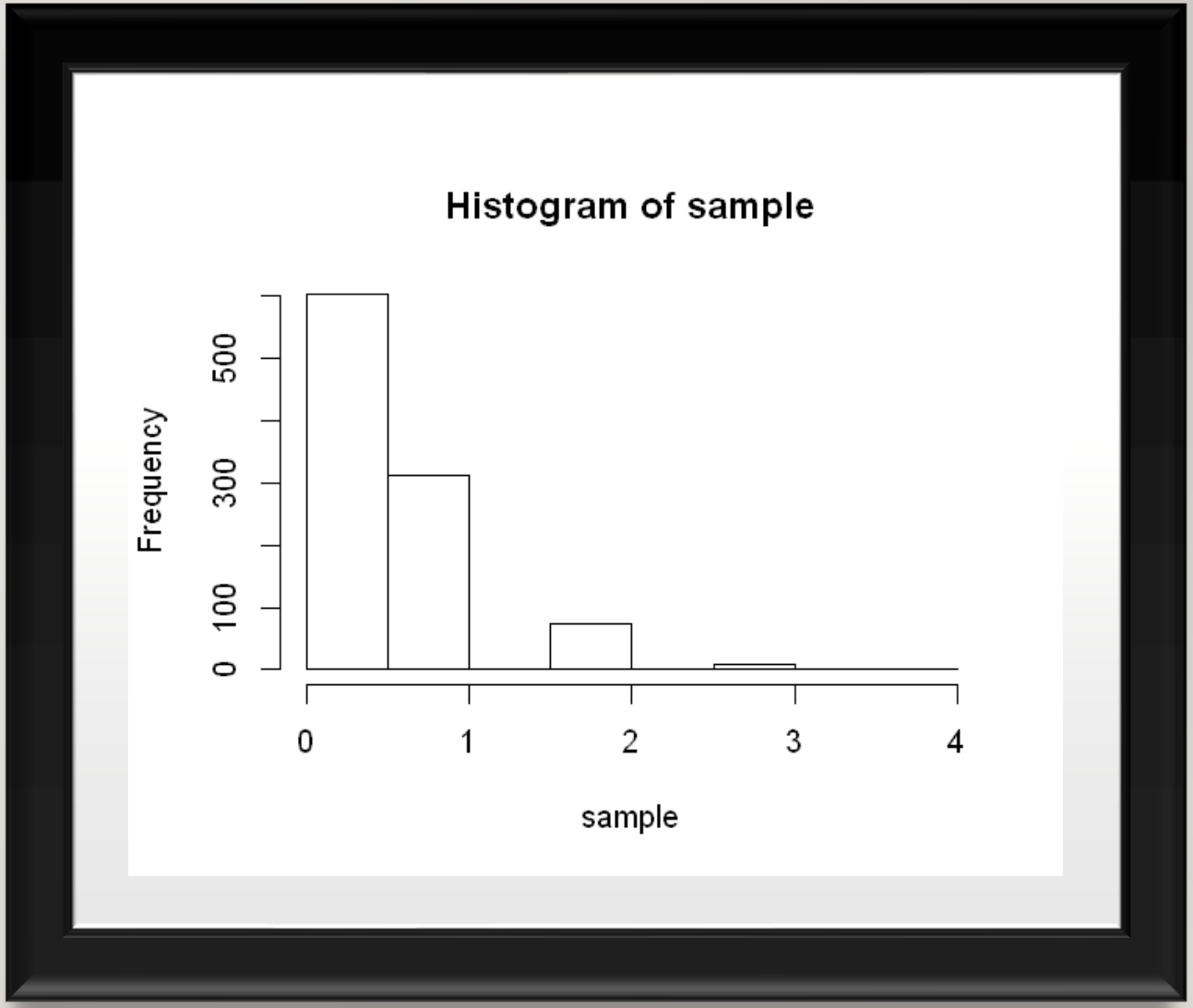
- $\nu \equiv Degrees\ of\ freedom$
- Degrees of freedom: A parameter that allows the t-distribution to adjust to different sample sizes, statistics, and number of groups.
- It is widely used as a refence basis for the distribution of sample means, differences between two sample means, regression parameters, and more.

BINOMIAL DISTRIBUTION

- Yes/No (binomial) outcomes lie at the heart of analytics since they are often the culmination of a decision or other process.
  - **Trial**: An event with a discrete outcome.
- The binomial distribution is the frequency distribution of the number of successes (x) in a given number of trials (n) with specified probability (p) of success in each trial.
- With large n, and provided p is not to close to 0 or 1, the binomial distribution can be approximated by the normal distribution. (**proof***)

BINOMIAL DISTRIBUTION

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad 0 \leq p \leq 1$$

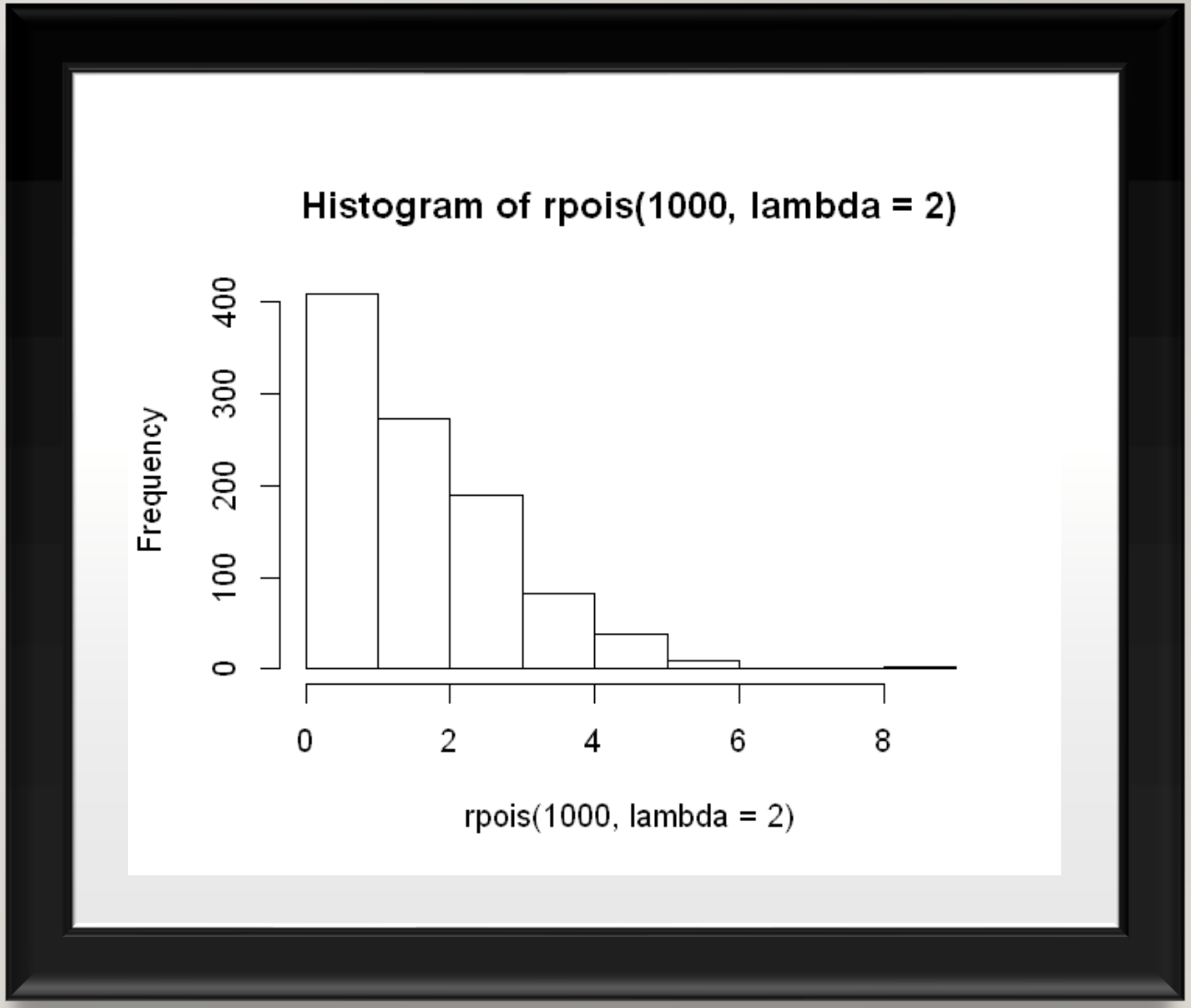


POISSON DISTRIBUTION

- The Poisson distribution tell us the distribution of events per unit of time or space when we sample many such units.
- “Internet traffic that arrives on a server in any 5-second period”
- “Number of car that cross a bump in any 5-minutes period”

POISSON DISTRIBUTION

$$f(k, \lambda) = \frac{e^{-\lambda} \lambda^k}{k!}$$

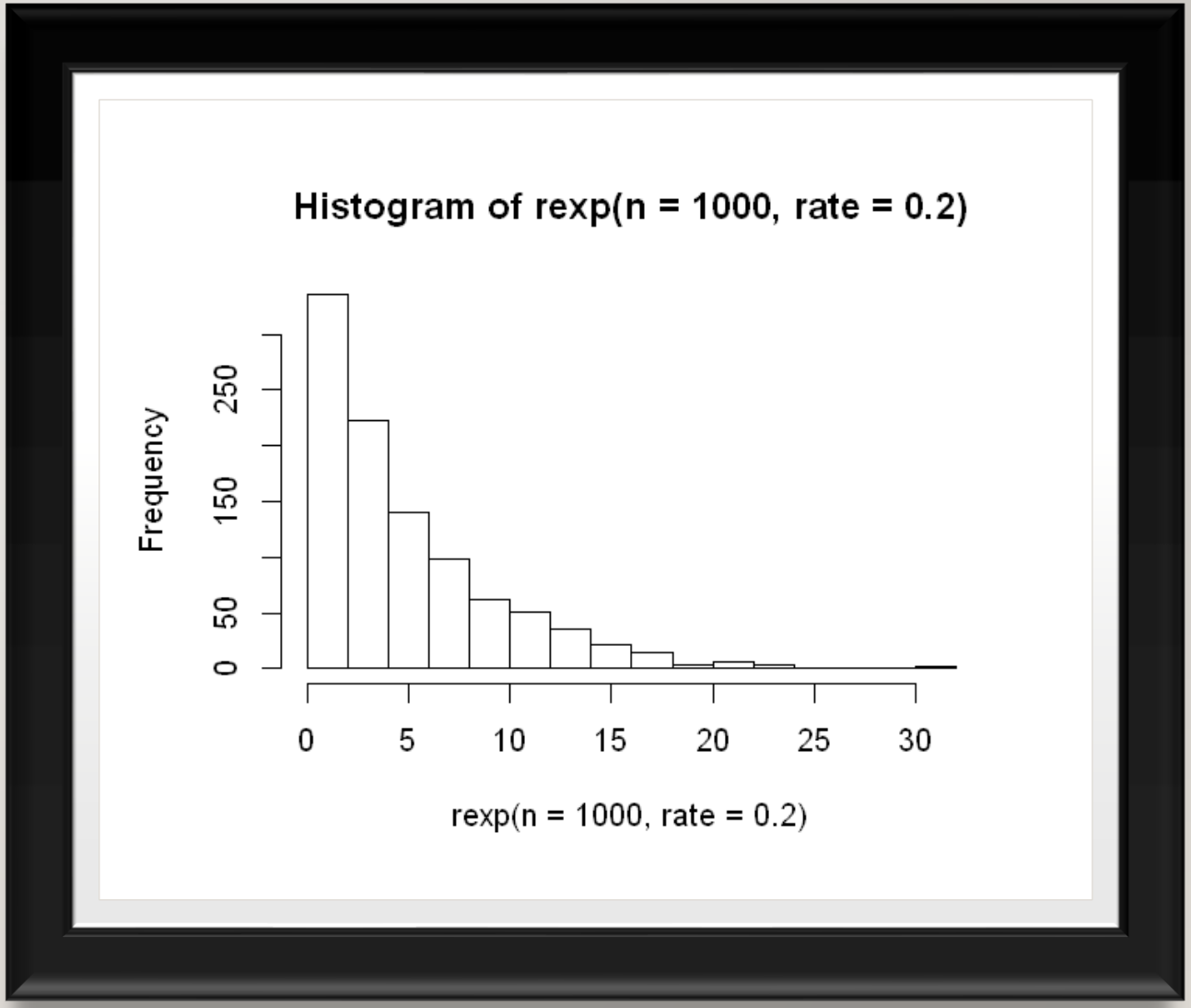


EXPONENTIAL DISTRIBUTION

- Models the distribution of the time between events.
- “Time between visits to a website or between cars arriving at a toll plaza”
- “Time required per service call due a product failure”

EXPONENTIAL DISTRIBUTION

$$f(x) = P(x) = \begin{cases} \lambda e^{-\lambda x} \\ 0 \end{cases}$$

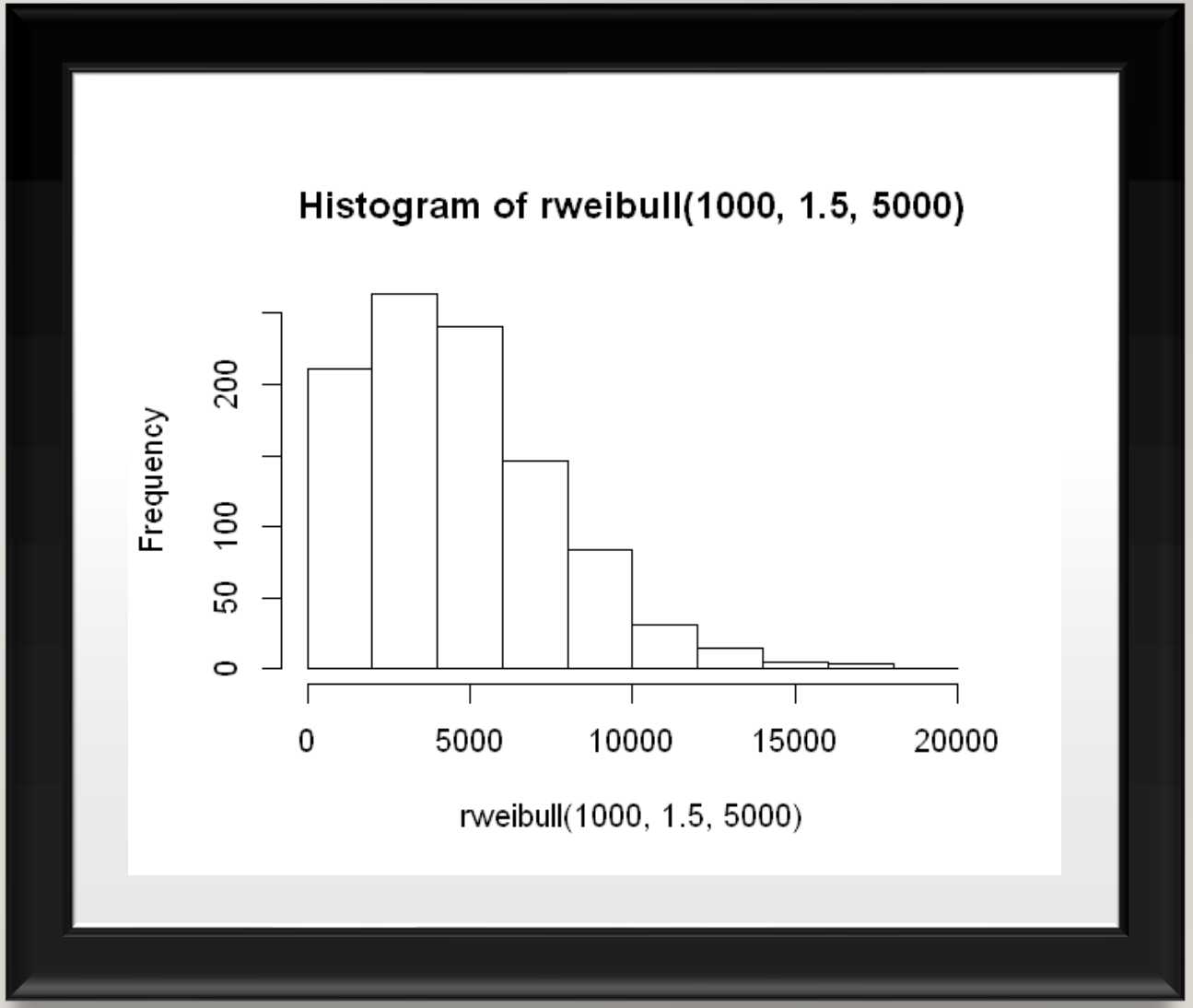


WEIBULL DISTRIBUTION

- The Weibull distribution is an extension of the exponential distribution, in which the event rate is allowed to change.
- Increasing probability of device failure, aircraft failure.

WEIBULL DISTRIBUTION

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} \\ 0 \end{cases}$$

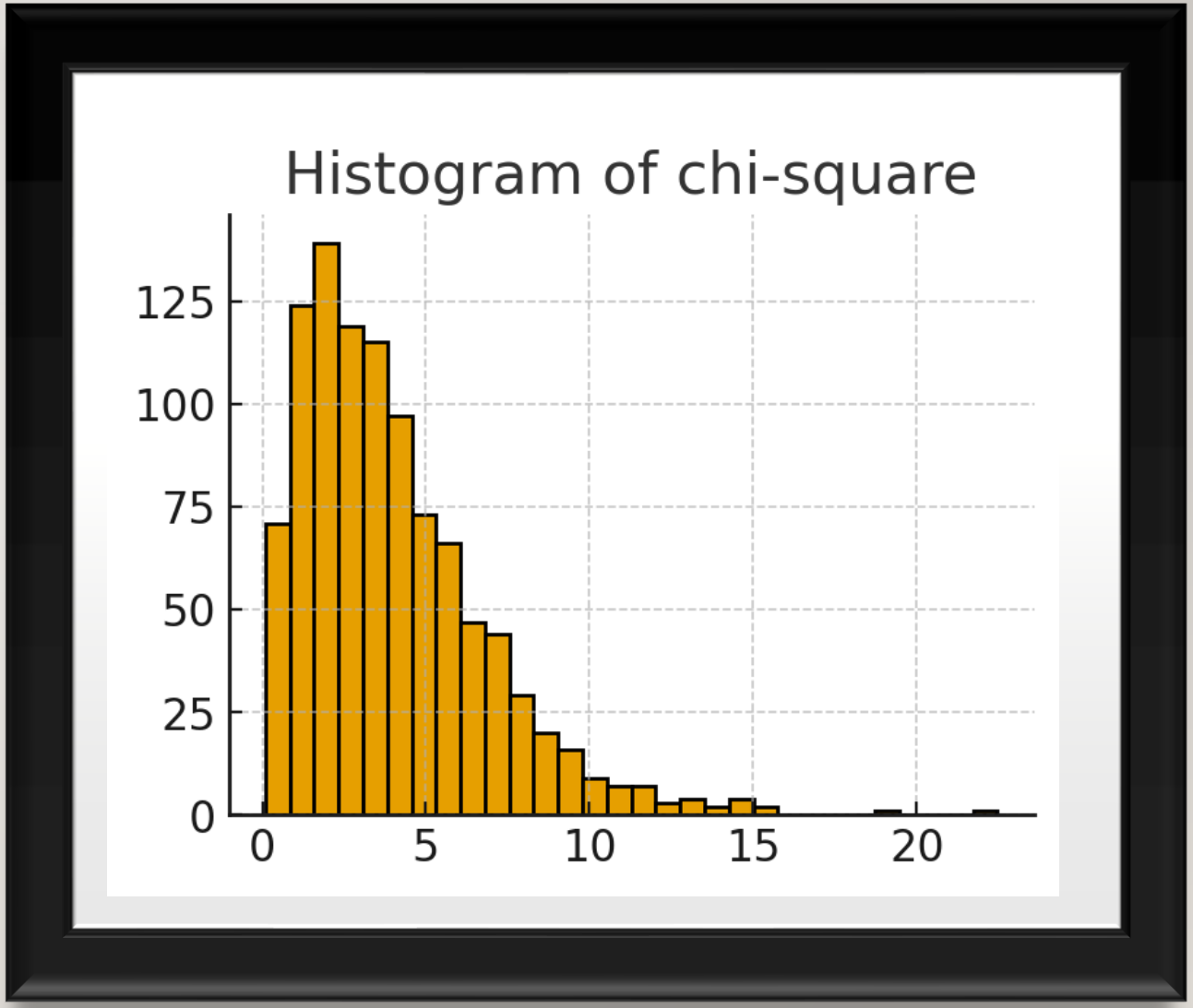


CHI-SQUARE DISTRIBUTION

- The Chi-Square distribution is widely used in hypothesis testing and confidence interval estimation for variance.
- It arises from the sum of the squares of independent standard normal variables.
- Applications:
  - Goodness of fit test for categorical data
  - Test of independence in contingency tables

CHI-SQUARE DISTRIBUTION

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2},\ x > 0$$

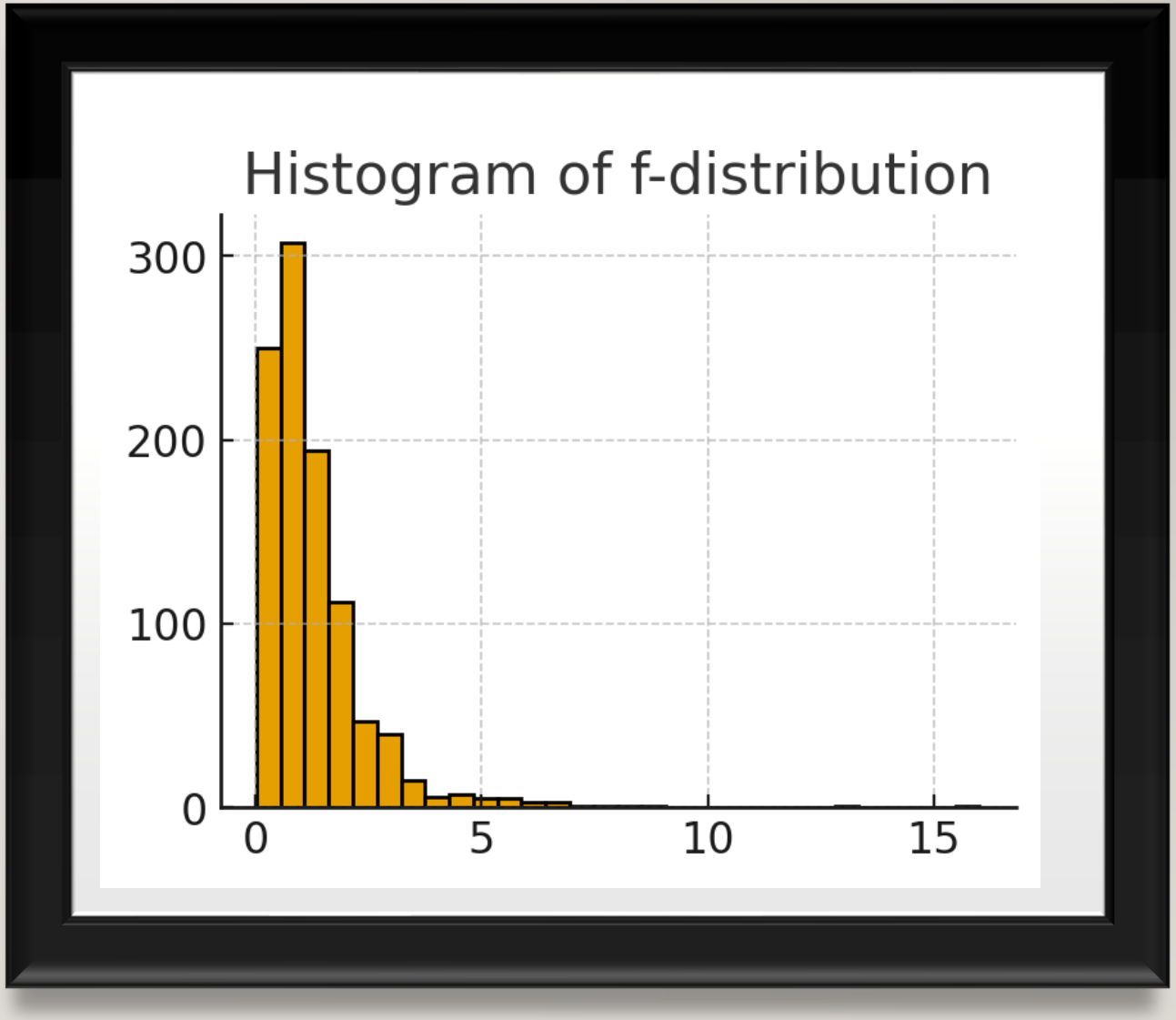


F DISTRIBUTION

- The F distribution is the ratio of two scaled chi-square distributions.
- It is commonly used to compare variances and in analysis of variance (ANOVA).
- Applications:
  - Testing if two populations have equal variances
  - ANOVA for comparing multiple group means

F DISTRIBUTION

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x B(d_1/2, d_2/2)}, \quad x > 0$$



SUMMARY

- Knowledge of various sampling and data generating distributions allows us to quantify potential errors in estimate that might be due to random variation.
- **Bootstrap** is an attractive "one size fits all" method to determine possible error in a sample estimates.