# ISE 599
# Special Topics applied predictive Analytics

Mayank Kejriwal

Research Assistant Professor/Research Lead

Department of Industrial and Systems Engineering

Information Sciences Institute

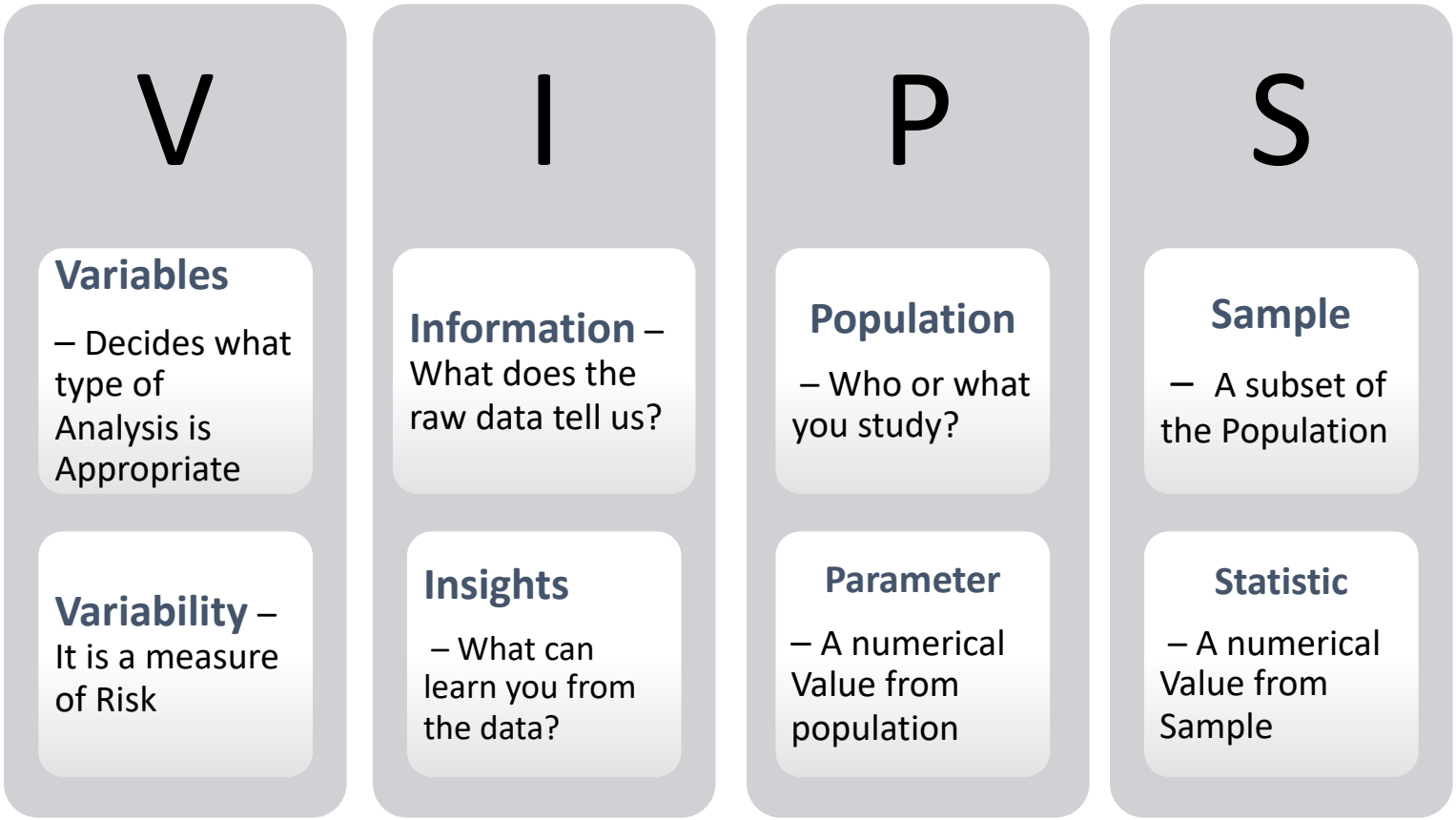
USC Viterbi School of Engineering

kejriwal@isi.edu

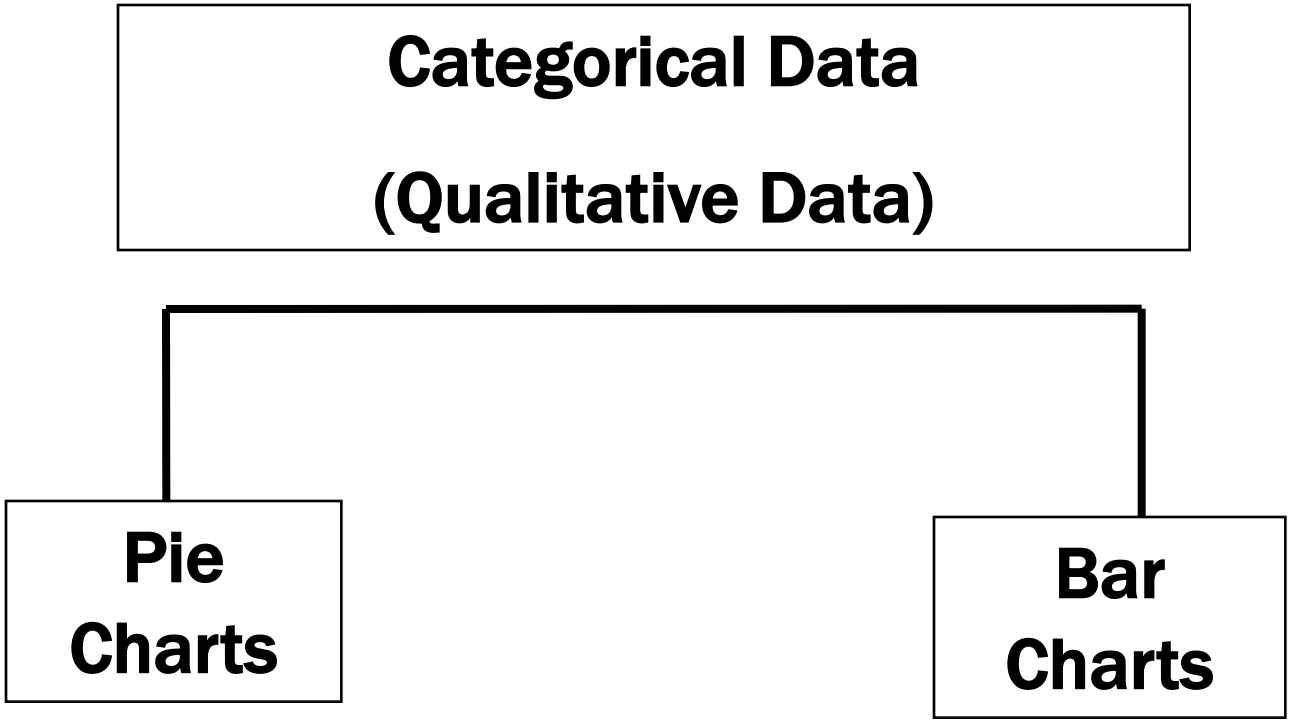
# 5-minute Review

# Analytics: some broad views

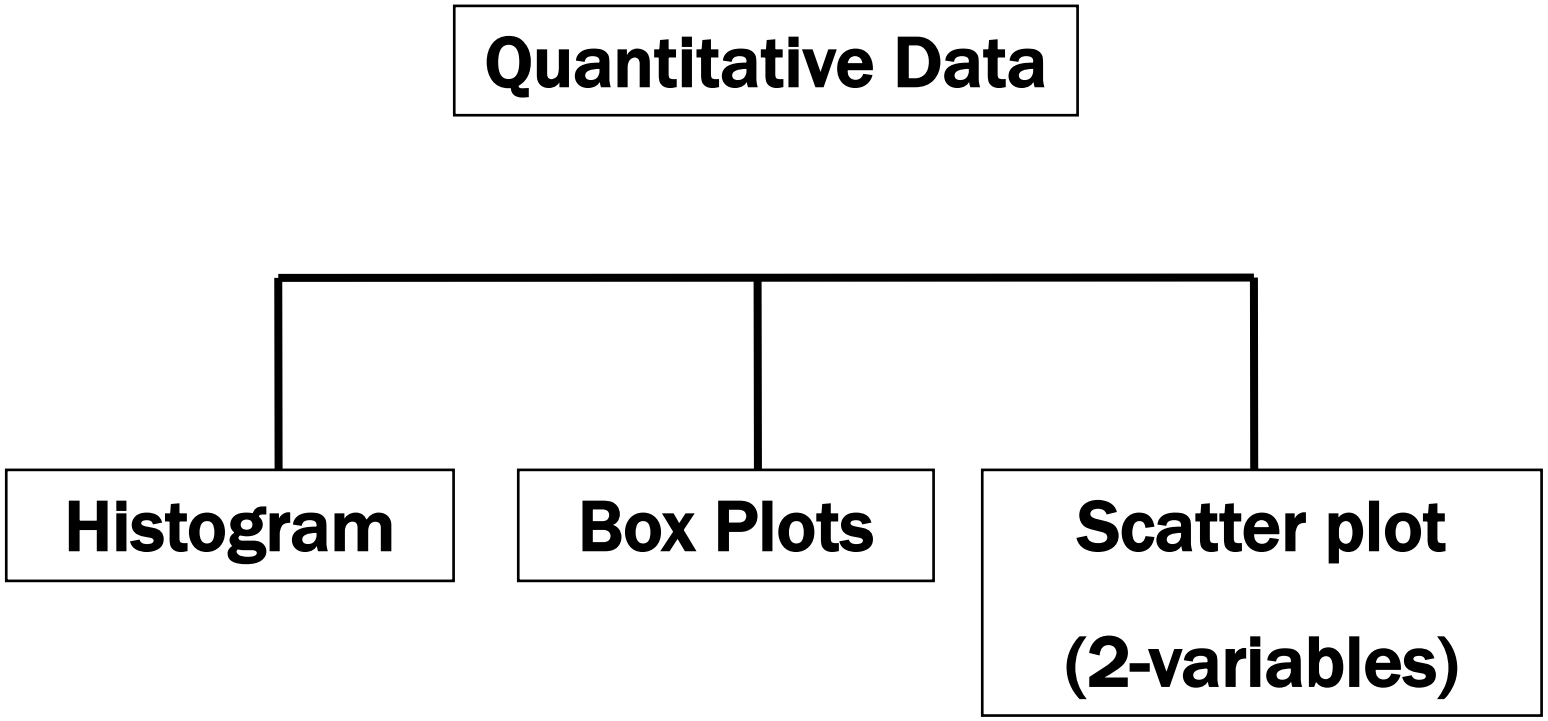
- From Oxford Dictionary:
  - the **systematic** computational analysis of **data or statistics**
  - information resulting from the **systematic** analysis of **data or statistics**
- From Techopedia:
  - Analytics is the **scientific process** of discovering and **communicating** the meaningful patterns which can be found in data. It is concerned with **turning raw data into insight for making better decisions**. Analytics relies on the application of statistics, computer programming, and operations research in order to quantify and gain insight to the meanings of data. It is especially useful in areas which record a lot of data or information.
- From 'Innovating with Analytics', MIT Sloan Management Review:
  - "…there is a strong correlation between driving competitive advantage and innovation with analytics and a company's effectiveness at managing the **information transformation cycle**, that is: **capturing data, analyzing information, aggregating and integrating data, using insights to guide future strategy and disseminating information and insights**."


V
Variables
– Decides what type of Analysis is Appropriate
Variability – It is a measure of Risk
I
Information – What does the raw data tell us?
Insights
– What can learn you from the data?
P
Population
– Who or what you study?
Parameter
– A numerical Value from population
S
Sample
– A subset of the Population
Statistic
– A numerical Value from Sample

# Graphing Categorical Data



# Graphing Quantitative Data



# Confounds

- Confounding occurs when the experimental controls do not allow the experimenter to **reasonably eliminate plausible alternative explanations** for an observed relationship between independent and dependent variables.

# Correlation vs. Causation

- The use of a controlled study is the most effective way of establishing causality between variables. In a controlled study, the sample or population is split in two, with both groups being comparable in almost every way. The two groups then receive different treatments, and the outcomes of each group are assessed.

# When making comparisons…

- …Make sure to choose a suitable frame of reference!
- Single biggest cause of misleading (sometimes intentional) statistics, especially in business and science

Lesson 1: Statistics (+visualizations) are the first line of attack in most applied analytics pipelines

# “There are three kinds of lies – lies, damned lies, and statistics”

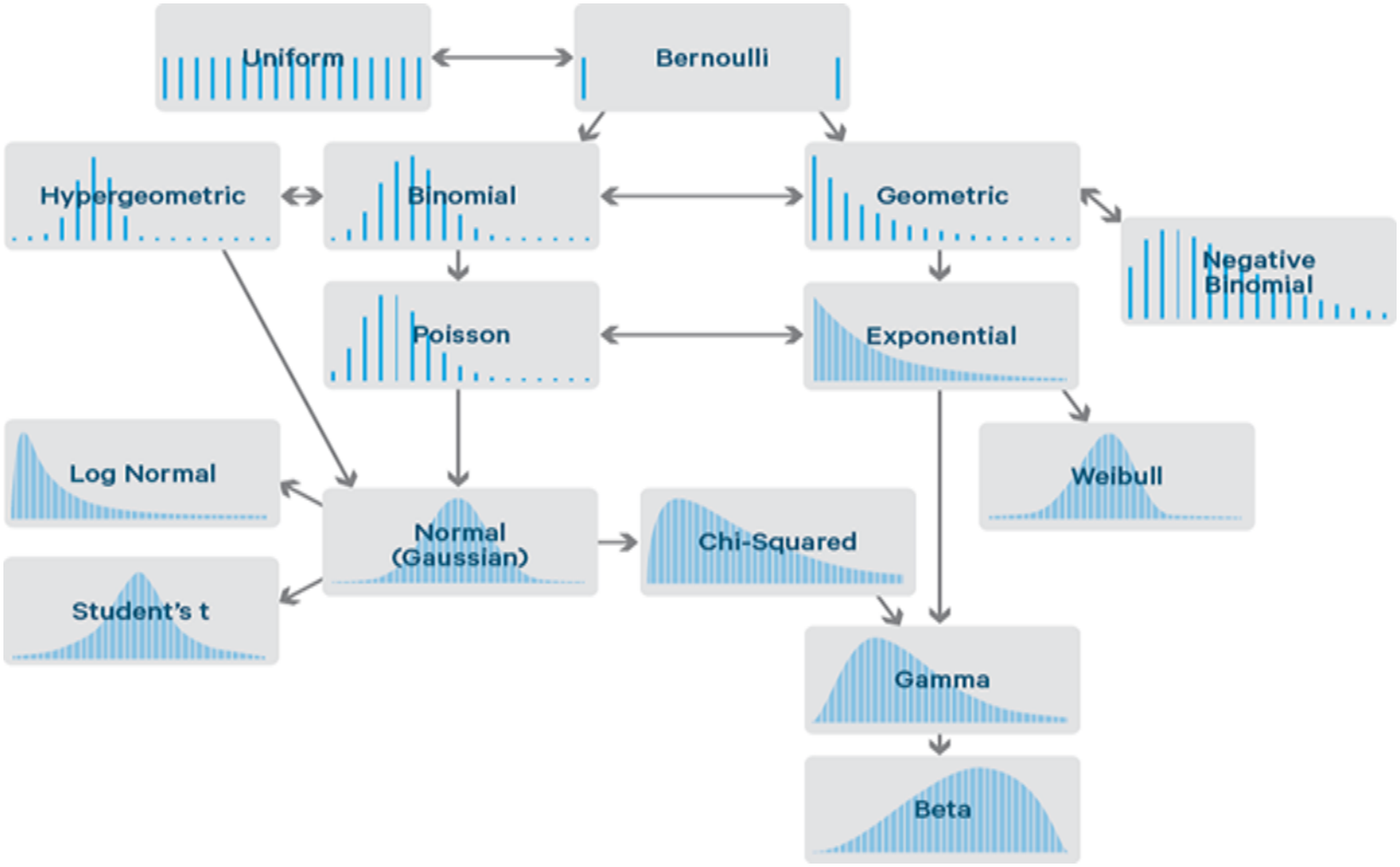
- To be an intelligent consumer of statistics, your first reflex must be to question the statistics you encounter
  - The more statistical knowledge and experience you have, the more adept you will be at doing the questioning
  - Like everything in life, skepticism should also be exercised in moderation
    - [Do not be like] Metrodorus of Chios: “None of us knows anything, not even this, whether we know or we do not know; nor do we know what ‘to not know’ or ‘to know’ are, nor on the whole, whether anything is or is not”

# Takeaways

- Think about statistics like scales in music: no one really ‘enjoys’ it, but without mastering it, you cannot be a good musician
- In the classroom, we state the problem in an obvious way but in the real world modeling the problem is half the battle!

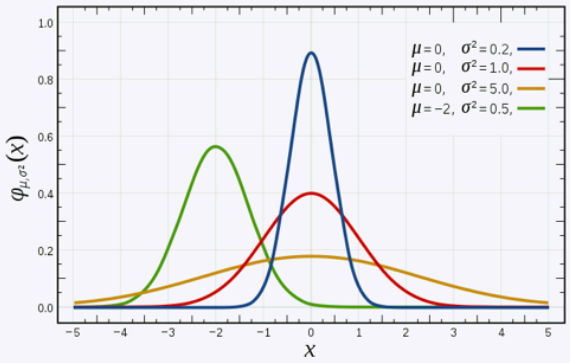
# probability

# What about probability?



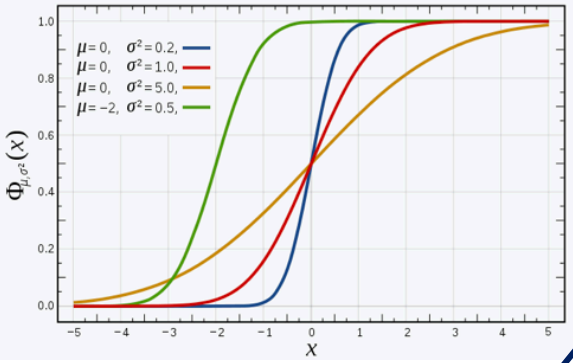
## Normal Distribution

Probability density function




The red curve is the *standard normal distribution*

Cumulative distribution function




$$\operatorname{erf}(z) \equiv \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt.$$

| | |
|---|---|
| **Notation** | $\mathcal{N}(\mu, \sigma^2)$ |
| **Parameters** | $\mu \in \mathbb{R}$ = mean (location)<br>$\sigma^2 > 0$ = variance (squared scale) |
| **Support** | $x \in \mathbb{R}$ |
| **PDF** | $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ |
| **CDF** | $\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$ |
| **Quantile** | $\mu + \sigma\sqrt{2}\,\operatorname{erf}^{-1}(2F-1)$ |
| **Mean** | $\mu$ |
| **Median** | $\mu$ |
| **Mode** | $\mu$ |
| **Variance** | $\sigma^2$ |
| **Skewness** | $0$ |
| **Ex. kurtosis** | $0$ |
| **Entropy** | $\frac{1}{2}\log(2\pi e\sigma^2)$ |
| **MGF** | $\exp(\mu t + \sigma^2 t^2/2)$ |
| **CF** | $\exp(i\mu t - \sigma^2 t^2/2)$ |
| **Fisher information** | $\mathcal{I}(\mu,\sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}\ \mathcal{I}(\mu,\sigma^2) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}$ |
| **Kullback-Leibler divergence** | $D_{\mathrm{KL}}(\mathcal{N}_0 \| \mathcal{N}_1) = \frac{1}{2}\{(\sigma_0/\sigma_1)^2 + \frac{(\mu_1-\mu_0)^2}{\sigma_1^2} - 1 + 2\ln\frac{\sigma_1}{\sigma_0}\}$ |

Acknowledgement: Wikipedia

# Let’s play a real-estate parlor game

- There is a list of houses whose selling prices are available and given to you (training data)
  - Assume you have some statistical software/calculator available on hand and the data is big enough for statistical generalization
  - No other info (sq. ft., no. of bathrooms etc.) available other than selling prices
- Iteratively (say over ‘ten’ trials):
  - A house is randomly sampled from a new list (that is *similar* to the old list) whose selling prices are hidden from us
  - Everyone has to estimate the selling price of the house (if you want to do some statistics, you have time)
  - To simplify matters, pick one selling price as your estimate for all trials
  - After the trials are over, a final score is calculated and a winner is declared
- How do you **win** the game?

# How is the score calculated?

- What's your estimate if the score is calculated as:
  - All or nothing (1 if you get the selling price exactly right, otherwise 0; doesn't matter if you're off by 1 USD or 1 million)?
  - Sum of deviations (sum over [true selling price – estimate])?
  - Sum of **squared** deviations?

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

Intuition?

| | |
|---|---|
| **Notation** | $\mathcal{N}(\mu, \sigma^2)$ |
| **Parameters** | $\mu \in \mathbb{R}$ = mean (location)<br>$\sigma^2 > 0$ = variance (squared scale) |
| **Support** | $x \in \mathbb{R}$ |
| **PDF** | $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ |
| **CDF** | $\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$ |
| **Quantile** | $\mu + \sigma\sqrt{2}\,\operatorname{erf}^{-1}(2F-1)$ |
| **Mean** | $\mu$ |
| **Median** | $\mu$ |
| **Mode** | $\mu$ |
| **Variance** | $\sigma^2$ |
| **Skewness** | $0$ |
| **Ex. kurtosis** | $0$ |
| **Entropy** | $\frac{1}{2}\log(2\pi e\sigma^2)$ |
| **MGF** | $\exp(\mu t + \sigma^2 t^2/2)$ |
| **CF** | $\exp(i\mu t - \sigma^2 t^2/2)$ |
| **Fisher information** | $\mathcal{I}(\mu,\sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}$ $\mathcal{I}(\mu,\sigma^2) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}$ |
| **Kullback-Leibler divergence** | $D_{\mathrm{KL}}(\mathcal{N}_0 \| \mathcal{N}_1) = \frac{1}{2}\{(\sigma_0/\sigma_1)^2 + \frac{(\mu_1-\mu_0)^2}{\sigma_1^2} - 1 + 2\ln\frac{\sigma_1}{\sigma_0}\}$ |

Acknowledgement: Wikipedia

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

$$\mathrm{Kurt}[X] = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$

Intuition?

- Excess Kurtosis is K[x]-3
- Mesokurtic, Leptokurtic and platykurtic

| | $x$ |
|---|---|
| **Notation** | $\mathcal{N}(\mu, \sigma^2)$ |
| **Parameters** | $\mu \in \mathbb{R}$ = mean (location)<br>$\sigma^2 > 0$ = variance (squared scale) |
| **Support** | $x \in \mathbb{R}$ |
| **PDF** | $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ |
| **CDF** | $\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$ |
| **Quantile** | $\mu + \sigma\sqrt{2}\,\operatorname{erf}^{-1}(2F-1)$ |
| **Mean** | $\mu$ |
| **Median** | $\mu$ |
| **Mode** | $\mu$ |
| **Variance** | $\sigma^2$ |
| **Skewness** | $0$ |
| **Ex. kurtosis** | $0$ |
| **Entropy** | $\frac{1}{2}\log(2\pi e\sigma^2)$ |
| **MGF** | $\exp(\mu t + \sigma^2 t^2/2)$ |
| **CF** | $\exp(i\mu t - \sigma^2 t^2/2)$ |
| **Fisher information** | $\mathcal{I}(\mu,\sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}$ $\mathcal{I}(\mu,\sigma^2) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}$ |
| **Kullback-Leibler divergence** | $D_{\mathrm{KL}}(\mathcal{N}_0 \| \mathcal{N}_1) = \frac{1}{2}\{(\sigma_0/\sigma_1)^2 + \frac{(\mu_1-\mu_0)^2}{\sigma_1^2} - 1 + 2\ln\frac{\sigma_1}{\sigma_0}\}$ |

Acknowledgement: Wikipedia

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

$$\operatorname{Kurt}[X] = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$

$$H(X) = -\int_{-\infty}^{\infty} p(x)\log p(x)\,dx.$$

| | |
|---|---|
| **Notation** | $\mathcal{N}(\mu, \sigma^2)$ |
| **Parameters** | $\mu \in \mathbb{R}$ = mean (location)<br>$\sigma^2 > 0$ = variance (squared scale) |
| **Support** | $x \in \mathbb{R}$ |
| **PDF** | $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ |
| **CDF** | $\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$ |
| **Quantile** | $\mu + \sigma\sqrt{2}\,\operatorname{erf}^{-1}(2F-1)$ |
| **Mean** | $\mu$ |
| **Median** | $\mu$ |
| **Mode** | $\mu$ |
| **Variance** | $\sigma^2$ |
| **Skewness** | $0$ |
| **Ex. kurtosis** | $0$ |
| **Entropy** | $\frac{1}{2}\log(2\pi e\sigma^2)$ |
| **MGF** | $\exp(\mu t + \sigma^2 t^2/2)$ |
| **CF** | $\exp(i\mu t - \sigma^2 t^2/2)$ |
| **Fisher information** | $\mathcal{I}(\mu,\sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}$ $\mathcal{I}(\mu,\sigma^2) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}$ |
| **Kullback-Leibler divergence** | $D_{\mathrm{KL}}(\mathcal{N}_0 \parallel \mathcal{N}_1) = \frac{1}{2}\{(\sigma_0/\sigma_1)^2 + \frac{(\mu_1-\mu_0)^2}{\sigma_1^2} - 1 + 2\ln\frac{\sigma_1}{\sigma_0}\}$ |

Acknowledgement: Wikipedia

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

$$\operatorname{Kurt}[X] = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$

$$H(X) = -\int_{-\infty}^{\infty} p(x)\log p(x)\,dx.$$

The moment-generating function of a random variable $X$ is

$$M_X(t) := \mathrm{E}\left[e^{tX}\right], \quad t \in \mathbb{R},$$

| | $X$ |
|---|---|
| **Notation** | $\mathcal{N}(\mu, \sigma^2)$ |
| **Parameters** | $\mu \in \mathbb{R}$ = mean (location)<br>$\sigma^2 > 0$ = variance (squared scale) |
| **Support** | $x \in \mathbb{R}$ |
| **PDF** | $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ |
| **CDF** | $\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$ |
| **Quantile** | $\mu + \sigma\sqrt{2}\,\operatorname{erf}^{-1}(2F-1)$ |
| **Mean** | $\mu$ |
| **Median** | $\mu$ |
| **Mode** | $\mu$ |
| **Variance** | $\sigma^2$ |
| **Skewness** | $0$ |
| **Ex. kurtosis** | $0$ |
| **Entropy** | $\frac{1}{2}\log(2\pi e\sigma^2)$ |
| **MGF** | $\exp(\mu t + \sigma^2 t^2/2)$ |
| **CF** | $\exp(i\mu t - \sigma^2 t^2/2)$ |
| **Fisher information** | $\mathcal{I}(\mu,\sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}$ $\mathcal{I}(\mu,\sigma^2) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}$ |
| **Kullback-Leibler divergence** | $D_{\mathrm{KL}}(\mathcal{N}_0 \| \mathcal{N}_1) = \frac{1}{2}\{(\sigma_0/\sigma_1)^2 + \frac{(\mu_1-\mu_0)^2}{\sigma_1^2} - 1 + 2\ln\frac{\sigma_1}{\sigma_0}\}$ |

Acknowledgement: Wikipedia

# Statistical Tests

## Null and Alternative Hypotheses

- Statistical hypothesis: claim about a parameter of a population.
- Null hypothesis ($H_0$): specifies a default course of action, preserves the **status quo.**
- Alternative hypothesis ($H_a$): contradicts the assertion of the null hypothesis,

- **$H_a$ : It is the Research Hypothesis →**
  - **What you want to test**
  - **The question you want to investigate**
  - **The statement for which you are collecting evidence**

# Example problem

The manager of a health maintenance organization has set as a target that the mean waiting time of non-emergency patients **will not exceed** 30 minutes (**Status Quo**). In spot checks, the manager finds the waiting times of 36 patients; the patients are selected randomly on different days. Assume that the **population standard deviation of waiting times is 10 minutes, the sample mean is 35 minutes (What is given? Status Quo or Research Hypothesis?)**

a. What is the relevant parameter to be tested?

*$\mu$ = mean waiting time of ALL non-emergency patients*

b. Formulate null and research hypotheses.

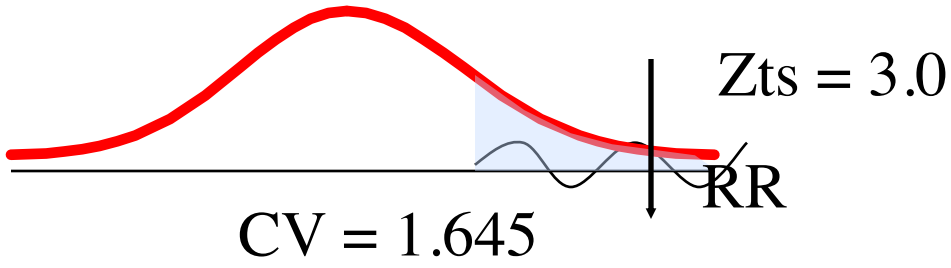
# Cont'd

c. State the test statistic and the rejection region corresponding to $\alpha = .05$.

$$\text{Test Statistic}: \qquad Z_{ts} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

$$Z_{ts} = \frac{35 - 30}{10 / \sqrt{36}} = \frac{5 * \sqrt{36}}{10} = 3.0$$

$$\text{Rejection Region}: \qquad \text{Reject } H_0 \text{ if } Z > 1.645$$



# Note 1: Central Limit Theorem (CLT)

- I will not state the CLT here, but I will say it is (typically) about the *sampling distributions of sample means*
  - Expected value of the sample mean does not change!
  - Variance declines over time, but it takes 30+ samples in practice before you can truly start assuming normality
- The theorem does not directly imply use of t-test etc. when doing significance testing (t-test assumes the data generator is normal to begin with, along with additional assumptions)
  - What do I do if data is non-normal? Use non-parametric tests!

# Note 2: Estimator vs. parameter

- Parameters are **not** random variables (they're fixed and are the statistical equivalent of 'constants'; hence, we use greek letters)
- Estimators are almost always (I've never seen an exception but make no claims that there aren't) random variables, and can be point estimates or intervals
  - Terminology is overloaded, but estimators are generally understood to be 'functions' that yield an estimate (e.g., given a set of (x,y) points, what is an estimator for 'b' in the eqn. y=bx+a)
  - Estimators could be biased, inconsistent (if sample size can vary), inefficient, have high MSE (or all of the above)
    - Use **bull's eye analogy** for intuition
  - Rare to find efficient and consistent estimators, but they do exist, especially in linear least-squares regression!
- **Question:** What does it 'mean' when I say my 95% confidence interval for a sample mean is [0.3,0.6]?

# Note 3: What if we have more than one null hypothesis?

- We can ‘exchange’ information between test statistics to gain power!
  - **Global hypothesis testing**: Simultaneously test all null hypotheses
  - **Multiple hypothesis testing:** separately test each null hypothesis

# Global hypothesis testing

We assume we have $n$ null hypotheses, $H_{0,i}$ with corresponding p-values $p_i$ obtained as if we were only testing hypothesis $i$ The global null hypothesis then is that all null hypotheses are true: $H_0 = \bigcap_{i=1}^{n} H_{0,i}$

Fisher's combination test:

(1) Compute T=−2∑log(p_i) based on the individual p-values, noting that under the null hypothesis, a p-value has a uniform distribution on [0,1]

(2) Under the global null hypothesis, T follows a chi-square distribution with 2n degrees of freedom, so we can decide to reject based on the value of T

# Bonferroni method (better known in the engineering/science community)

Reject $H_0$ if $\min_i p_i \leq \alpha/n$

"The way to get started is to quit talking and begin doing."
- Walt Disney