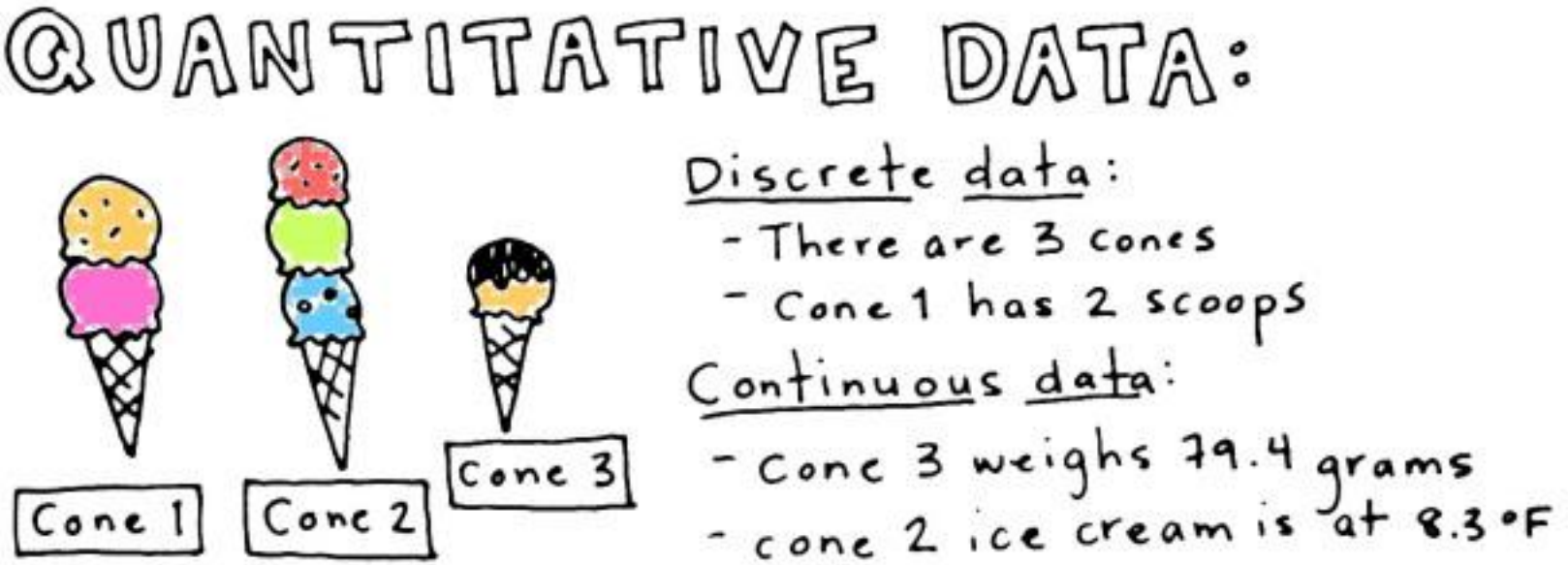
Exploratory Data Analysis

Data – It's numeric*

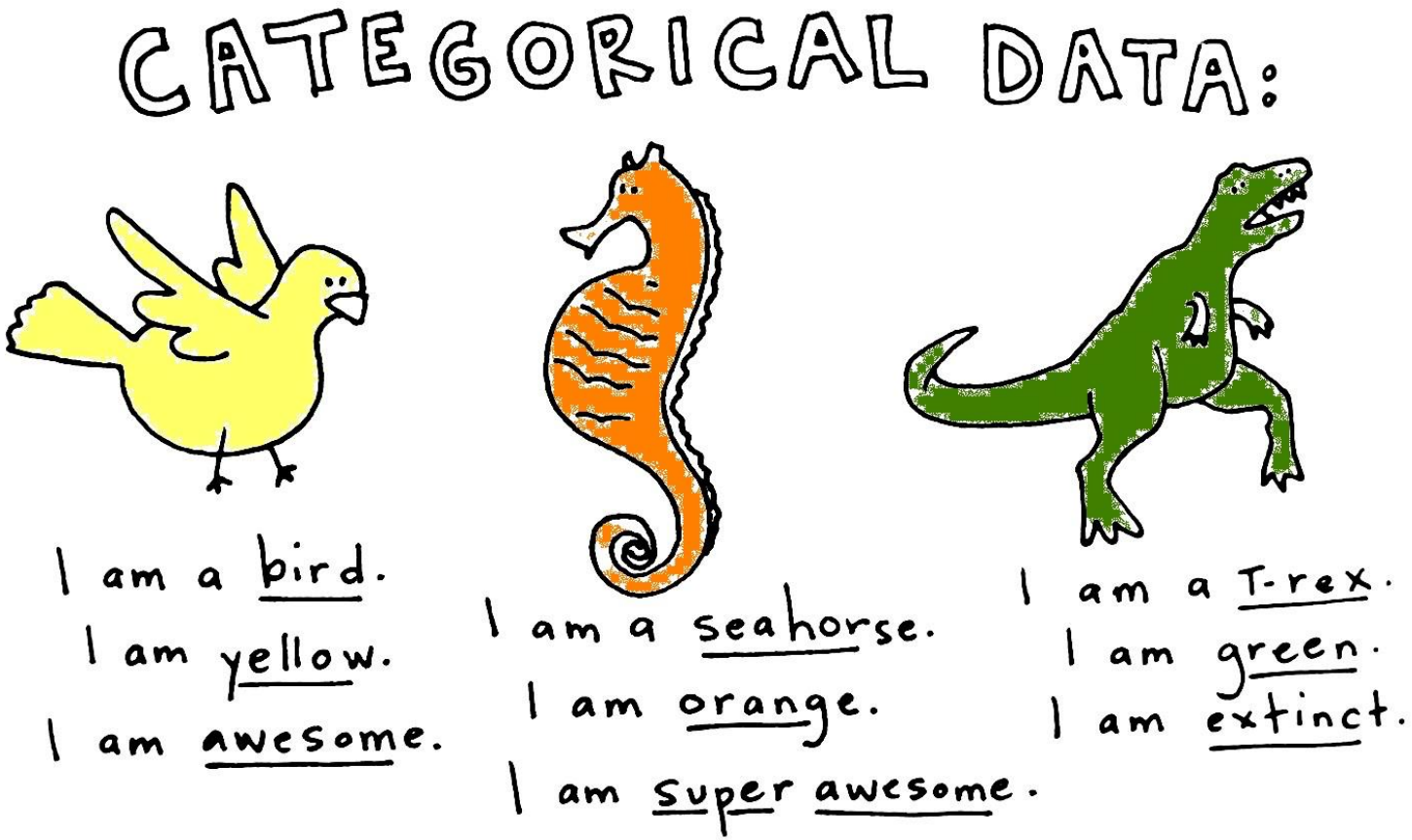


Discrete variables – only a few possible values and no in-between values

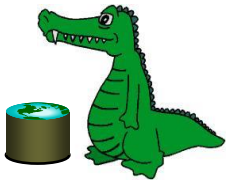
Continuous variables – several possible values and in-between values

*http://statsthewayilikeit.com/

Data – It's descriptive*

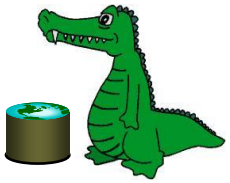


*http://statsthewayilikeit.com/



Random Variable*

- A random variable is a variable whose realization is determined by chance
  - Rolling a die
- One can describe a random variable by its expected value. It is the sum of all possible realizations weighted by their probabilities
- The expected value is similar to an average, but with an important difference: the average is computed when you already have realizations, while the expected value is computed before you have realizations



Mean, Variance, Standard Deviation

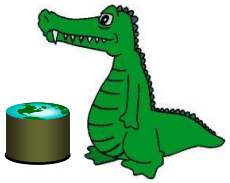
The mean is the arithmetic average of the observations.

The variance is a measure of spread of the individual observations from the sample mean:

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

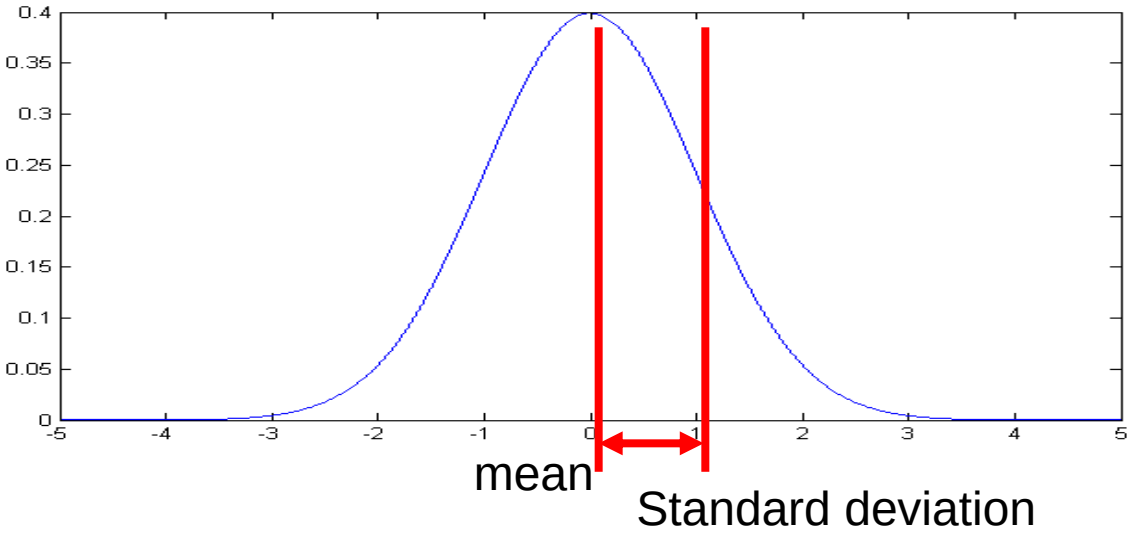
The standard deviation is the square root of variance.

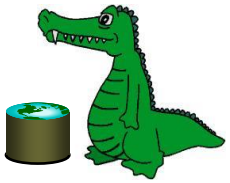
The probability density function (pdf) covers an area representing the probability of realizations of the underlying values



Normal Distribution

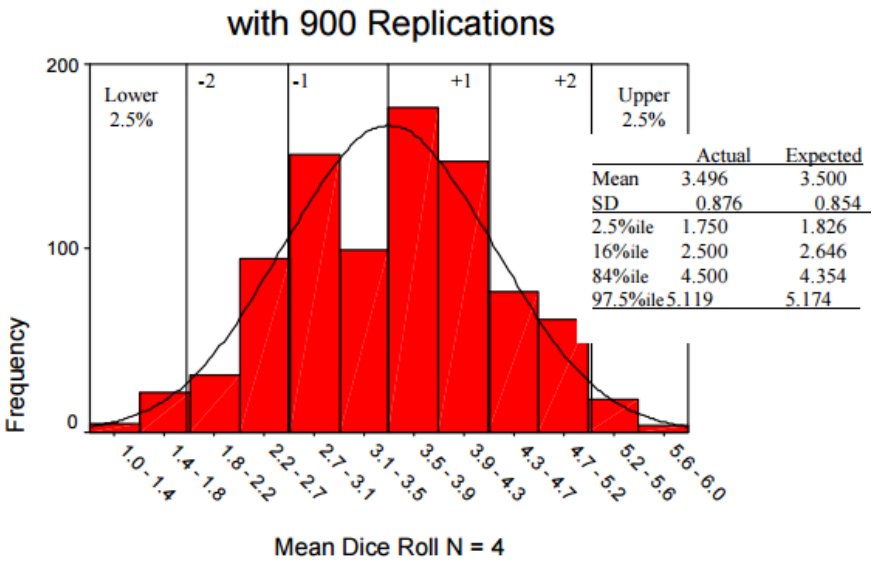
- Symmetric about the mean
- Mean, median and mode are the same
- Defined by mean (μ) and standard deviation (σ)
- Total area under the normal curve = 1 (pdf)

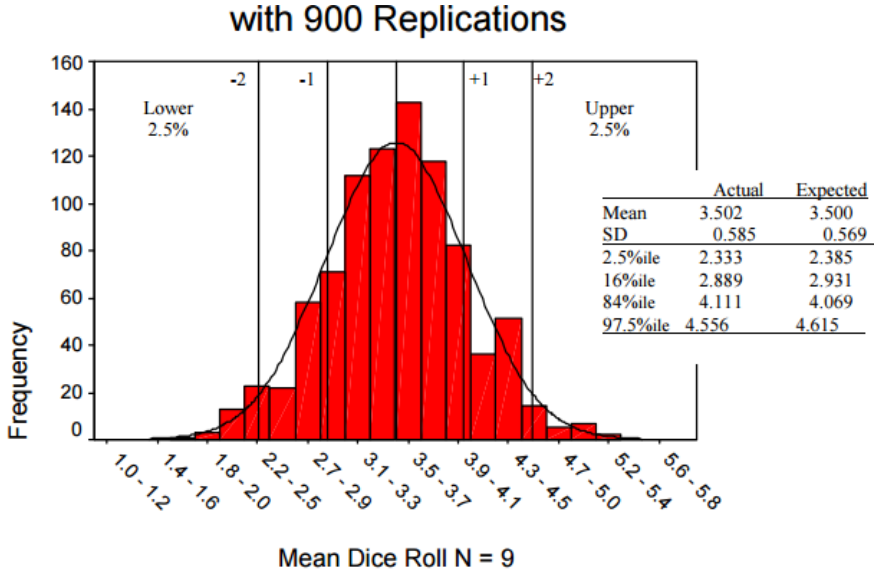




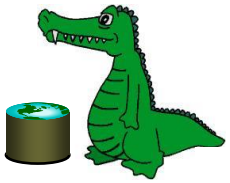
Central Limit Theorem

The distribution of the mean of a set of N identically-distributed random variables approaches a normal distribution as $N \rightarrow \infty$.

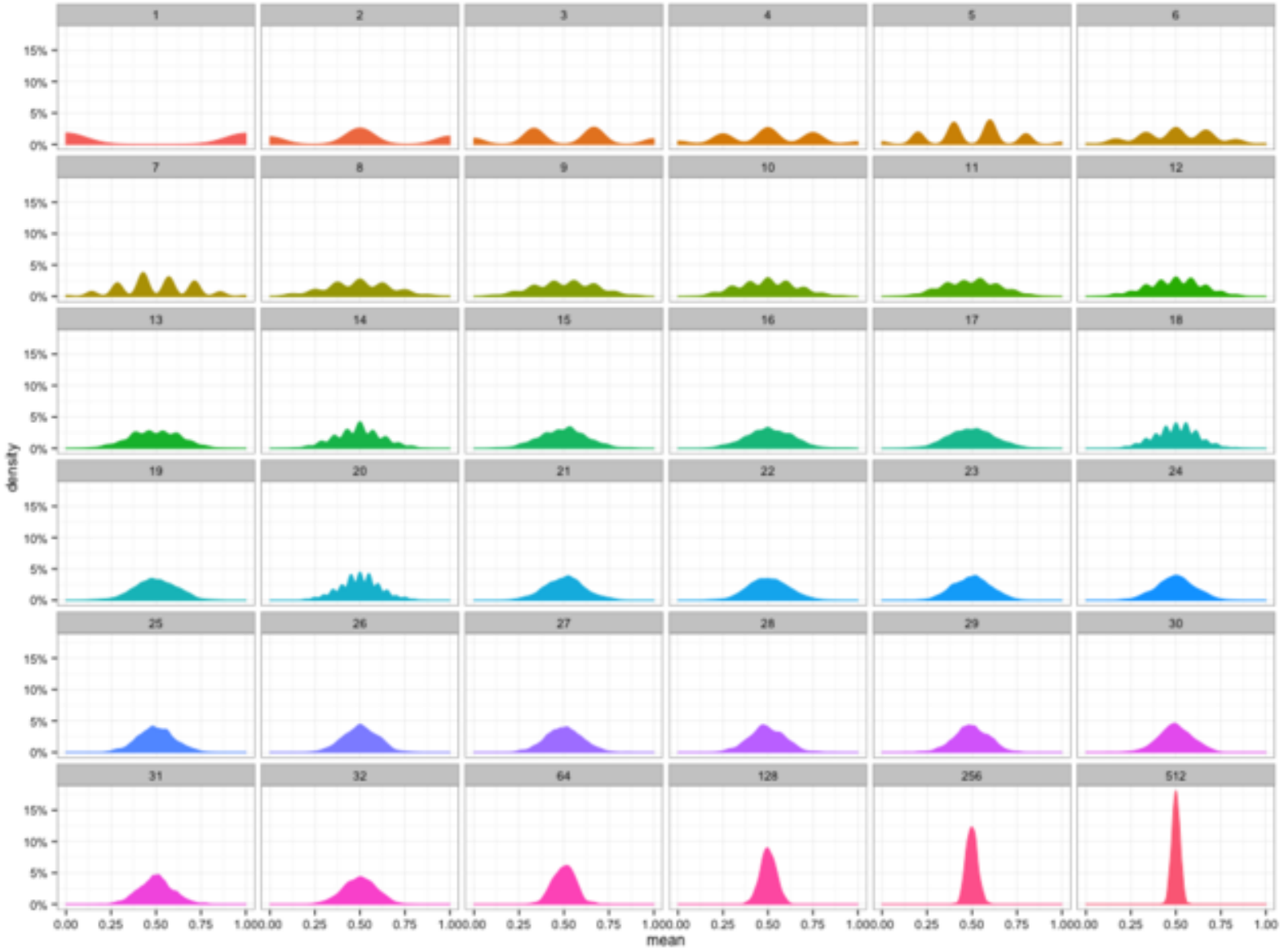




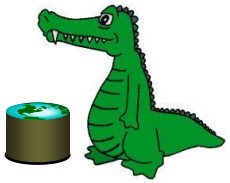
http://www.soph.uab.edu/ssg/files/People/MBeasley/Courses/CLT.pdf



Central Limit Theorem



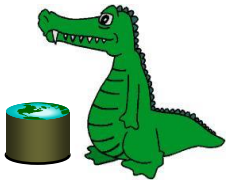
From the Wikipedia



Normality Assumption and CLT

All parametric statistical tests (e.g. t-test and ANOVA) assume normally-distributed data, but depend on sample mean and variance measures of the data.

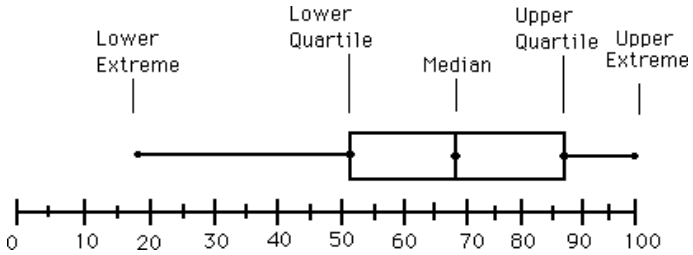
But, they work reasonably well for data that are not normally distributed as long as the samples are not too small (because of CLT).



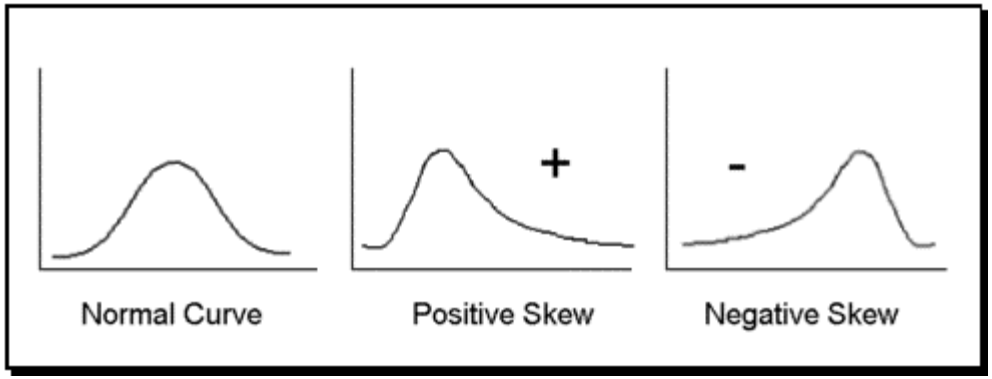
Normally distributed?

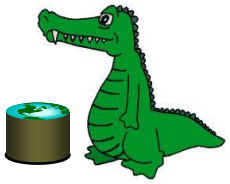
Many statistical tools, including mean and variance, t-test, ANOVA etc. assume **data** are **normally distributed**.

Very often this is not true. The box-and-whisker plot gives a good clue



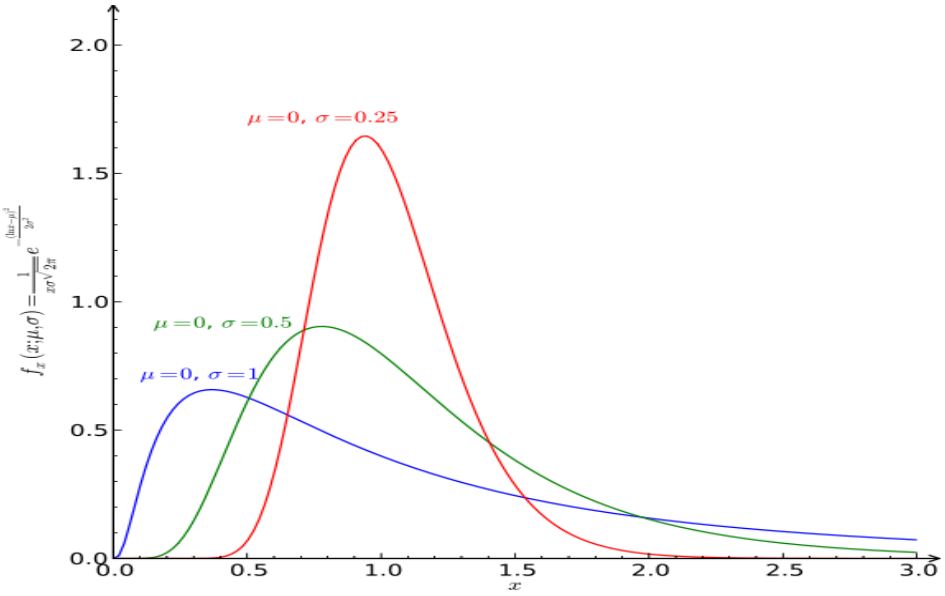
Whenever its asymmetric, the data cannot be normal. The histogram gives even more information

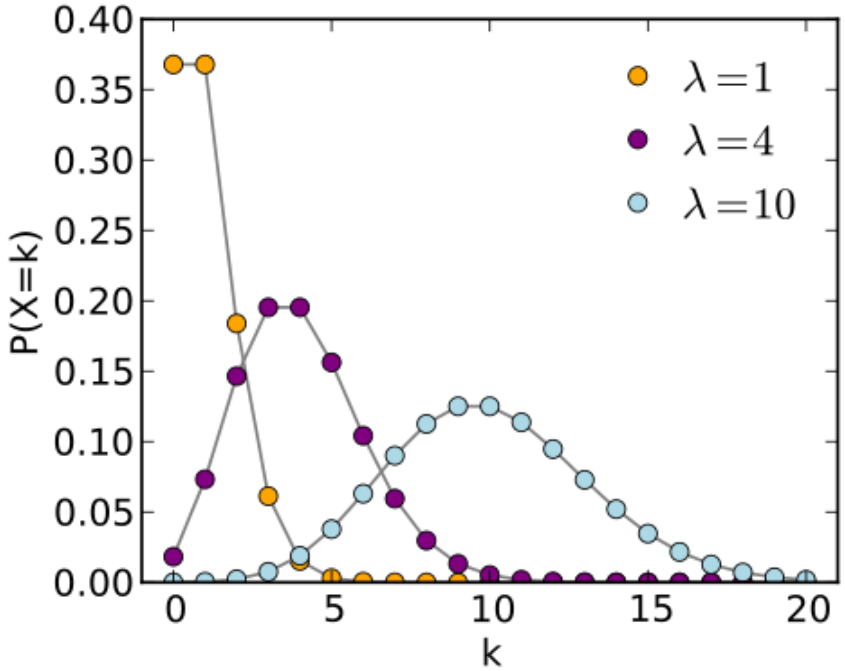




Correcting distributions

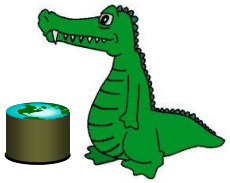
If X satisfies a log-normal distribution, $Y = \log(X)$ has a normal dist.





If X is a Poisson* with mean λ and sdev. $\sqrt{\lambda}$, then $\sqrt{X}$ is approximately normally distributed with sdev. 1 with $\lambda > 10$

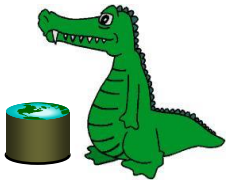
* See Additional Reading material



Inference

By inference, we mean the research of the values of the parameters given some data

- **Estimation**: use the data to estimate the parameters
- **Hypothesis Testing**: guess a value for the parameters and ask the data whether this value is true



Hypothesis Testing: Motivation

Suppose, for the past year, the mean of the monthly energy cost for families was $260 p.m.

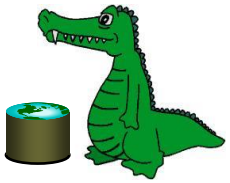
Determine whether the mean has changed, for the current year.

One solution: generate a **random sample** of 25 families and record energy costs for the current year.

Descriptive Statistics: Energy Cost

Variable	Total Count	Mean	SE Mean	StDev
Energy Cost	25	330.6	30.8	154.2

*http://blog.minitab.com

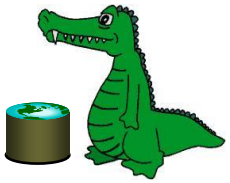


Hypothesis Testing: Motivation

Even though our *sample* mean is 330.6, the population mean could still be 260 due to **sampling error**.

Hypothesis testing can assess the likelihood of this possibility!

*http://blog.minitab.com

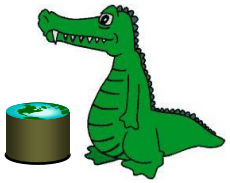


Null & Alternate Hypothesis

The **null hypothesis (H_0)**: The population mean (330.6) equals the hypothesized mean (260).

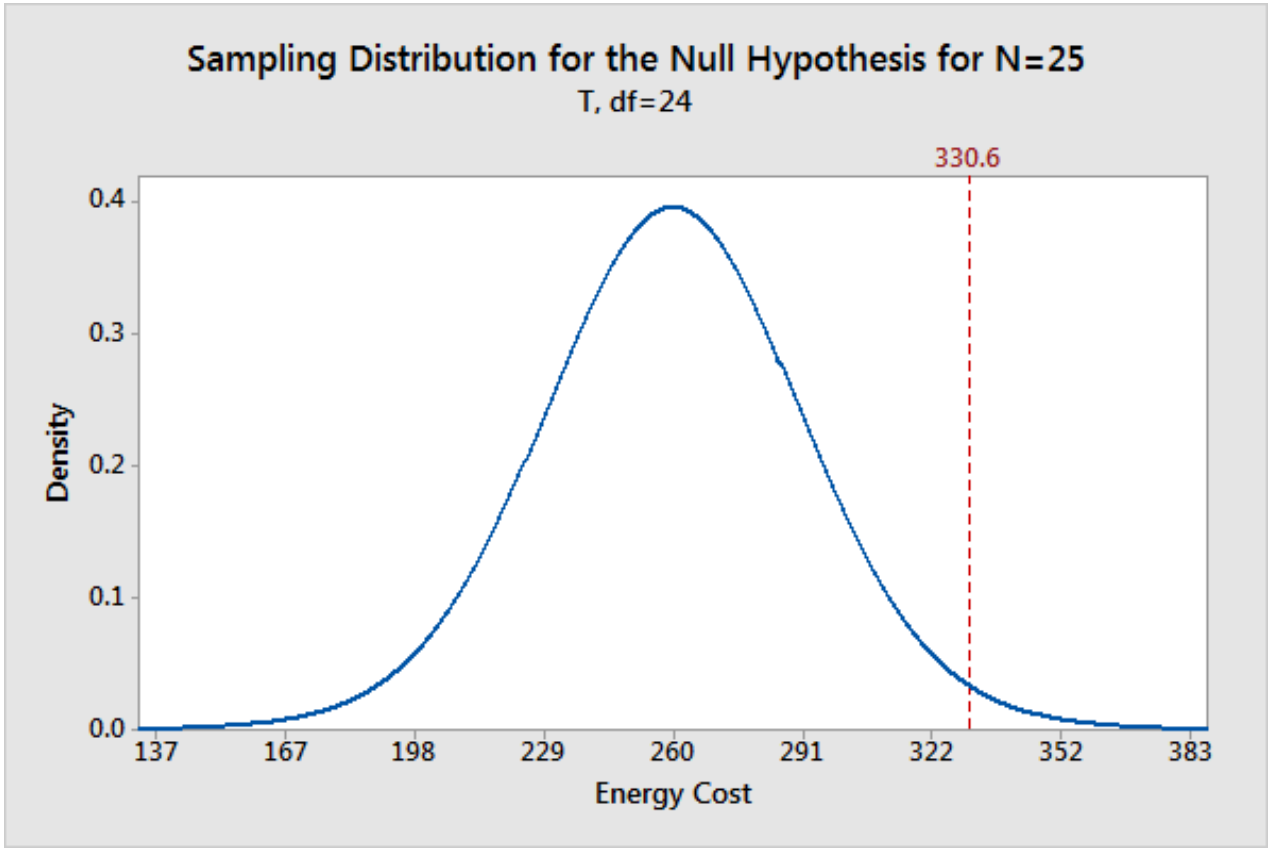
The **alternative hypothesis (H_a)**: The population mean (330.6) differs from the hypothesized mean (260).

A **sampling distribution** is the distribution of a **statistic**, such as the mean, that is obtained by repeatedly drawing a large number of samples from a specific population.

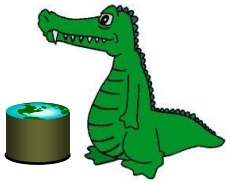


Hypothesis Testing

Goal is to determine whether our sample mean (330.6) is significantly different from the null hypothesis mean (260)

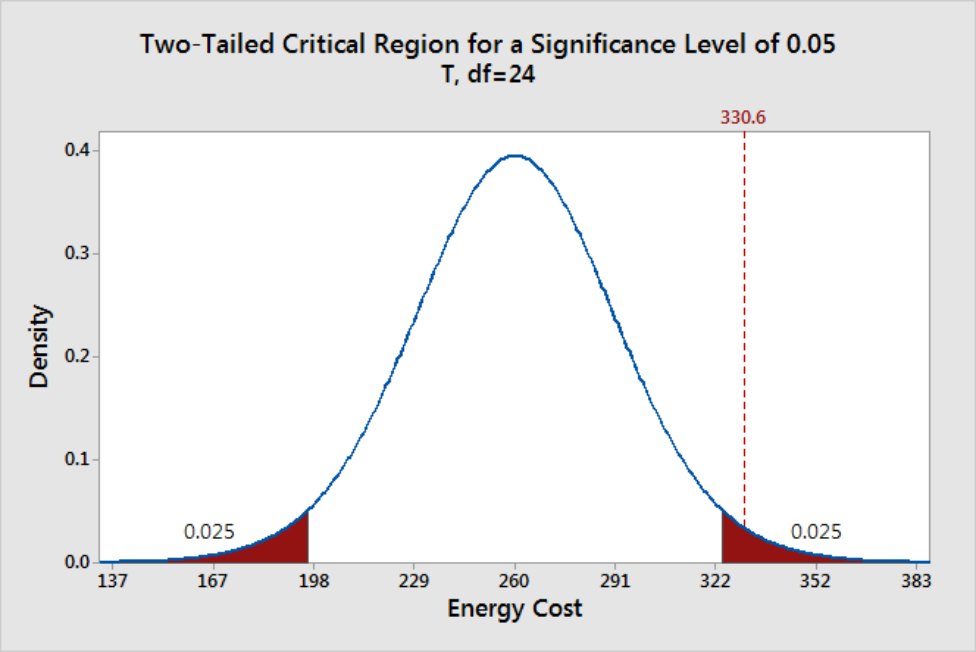


*http://blog.minitab.com



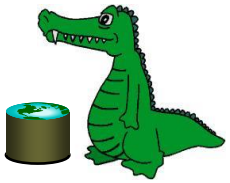
Significance Level α*

A significant level α defines a **critical region**, which indicates how far away our sample mean must be from the null hypothesis mean



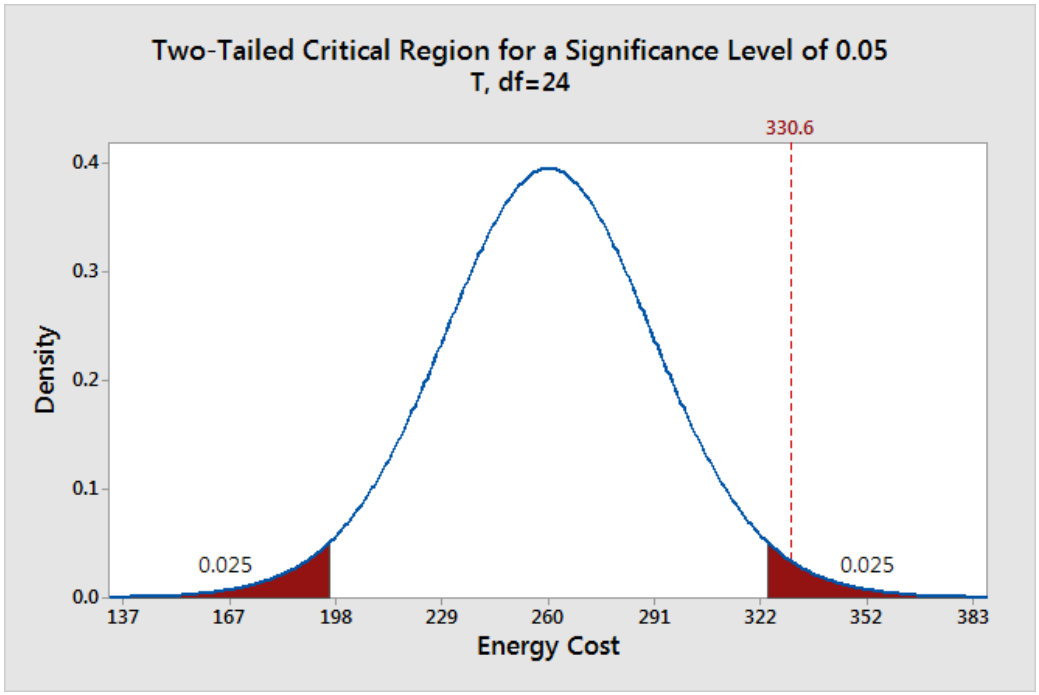
Our sample mean (330.6) falls within the critical region, which indicates it is statistically significant at the $\alpha = .05$ level

*http://blog.minitab.com



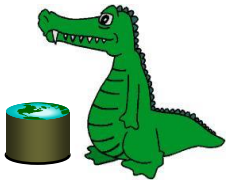
Significance Level α*

α is the probability of rejecting the null hypothesis when it is true, i.e., **Type I Error**



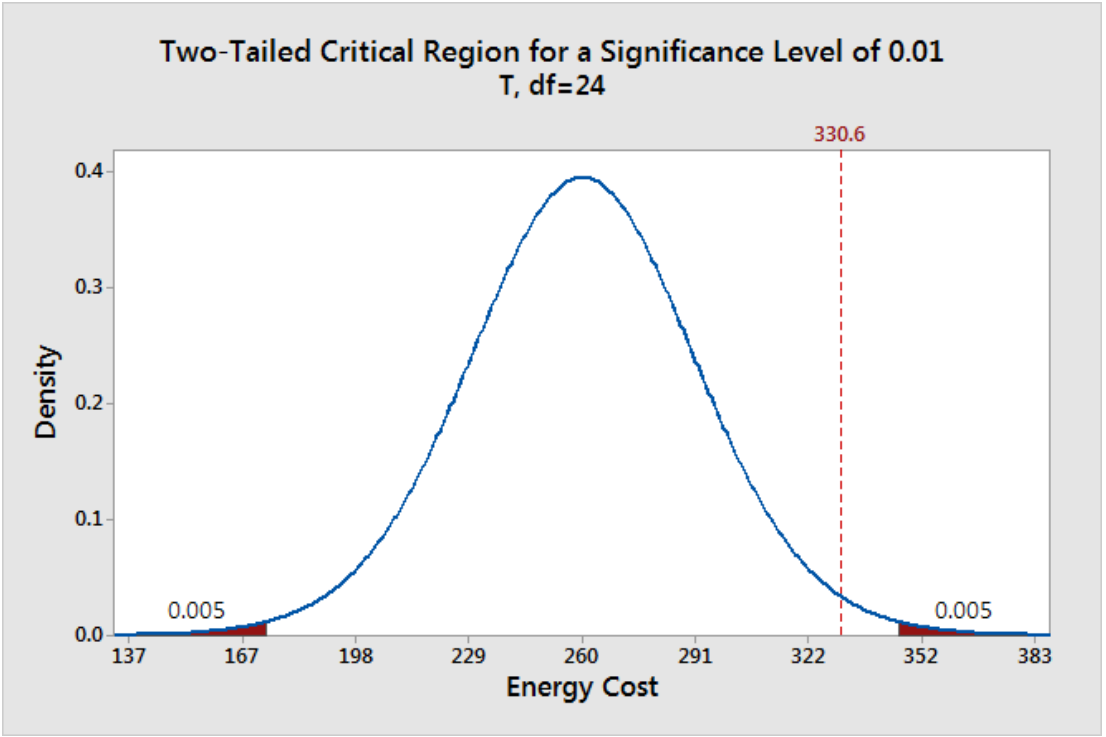
How far out do we draw the line for the critical region?

*http://blog.minitab.com



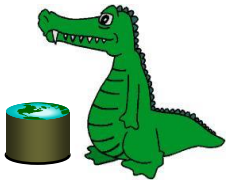
Significance Level α*

Another common $\alpha = .01$: our sample mean does not fall within the **critical region**



You needs to choose the significance level before you begin your study: Common $\boldsymbol{\alpha}$-values are .01, .05, etc.

*http://blog.minitab.com



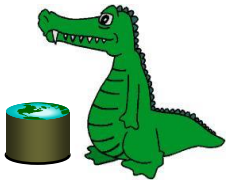
What are p-values

p-value represents the probability of obtaining an effect at least as extreme as the one in your sample data, assuming the truth of the null hypothesis.

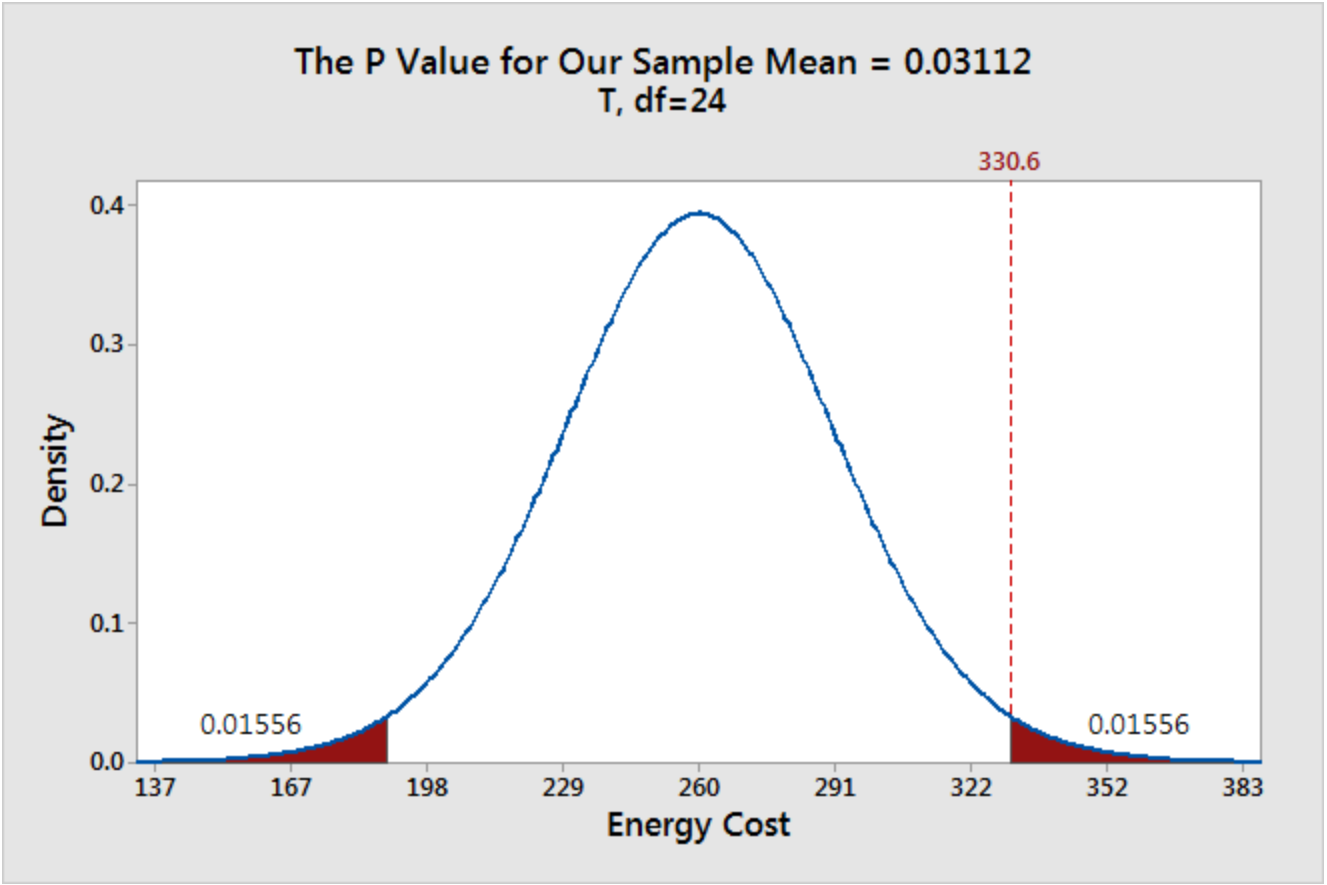
To compute the p-value for our example:

- Compute the distance between our sample mean and the null hypothesis value (330.6 - 260 = 70.6).
- Graph the probability of obtaining a sample mean that is at least as extreme in both tails of the null hypothesis distribution (i.e. 260 +/- 70.6).

*http://blog.minitab.com

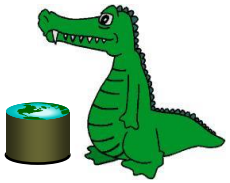


What are p-values



If p-value $\leq$ α, we reject the null hypothesis.

*http://blog.minitab.com

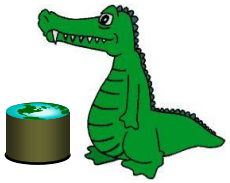


What are p-values

They measure how compatible your data are with the null hypothesis.

- High p-values: data are likely aligned with H_0
- Low p-values: data are unlikely aligned with H_0

But, they don't measure support for the alternative hypothesis.

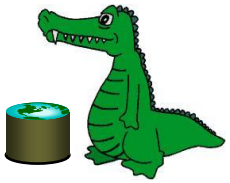


Three important tests

- **T-test:** compare two groups or two interventions on one group.

- **CHI-squared (Fisher's test):** Compare the counts in a "contingency table".

- **ANOVA:** Compare means in more than two different group of individuals.



One-Sample T-test

Checks the evidence that the mean (μ) of a sample $\neq \mu_0$.

Typical **null hypothesis** and **alternate hypothesis** are:

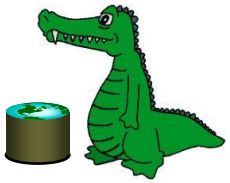
$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

One can also use this test for **one group** of individuals in **two conditions** (before vs. after): use the difference of the two measurements for each person

In this test, we compute the **test statistic** as

$$t = \frac{\bar{X}}{\bar{\sigma}}$$

where $\bar{X}$ is the sample mean and $\bar{\sigma}$ is the sample standard deviation.



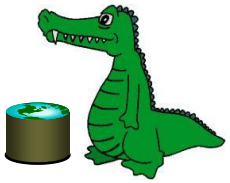
Two sample T-test

Used to compare the means from exactly **TWO** groups, such as the **control** group vs. the **experimental** group

$$H_0: \mu(X_1) = \mu(X_2)$$
$$H_a: \mu(X_1) \neq \mu(X_2)$$

Suppose there are **two samples** X_1 and X_2. A t-statistic is constructed from their sample means and sample standard deviations:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} \qquad s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$



A/B Testing*

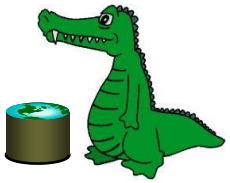
An example problem:

- Test a website landing page that has a signup form
- Test various layouts to try and maximize the "conversion rate", i.e., the percentage of people who sign up

Setup: a **control** treatment and 3 **experimental** treatments A, B, C

Note: It's similar to paired t-test

*http://20bits.com/article/statistical-analysis-and-ab-testing



Synthetic data

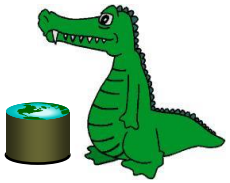
Goal: increase the landing page conversion rate by at least 20%

Project X Landing Page

Treatment	Visitors Treated	Visitors Registered	Conversion Rate
Control	182	35	19.23%
Treatment A	180	45	25.00%
Treatment B	189	28	14.81%
Treatment C	188	61	32.45%

Treatment C is "good enough”, but can you describe the goodness for e.g. with 95% confidence interval?

*http://20bits.com/article/statistical-analysis-and-ab-testing



The Statistics

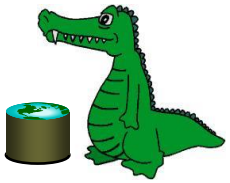
The **null hypothesis**

$$H_0: p - p_c \leq 0$$

p_c is the conversion rate of the control and p is the conversion rate of one of our experiments

The **alternative hypothesis** is that the experimental page has a higher conversion rate.

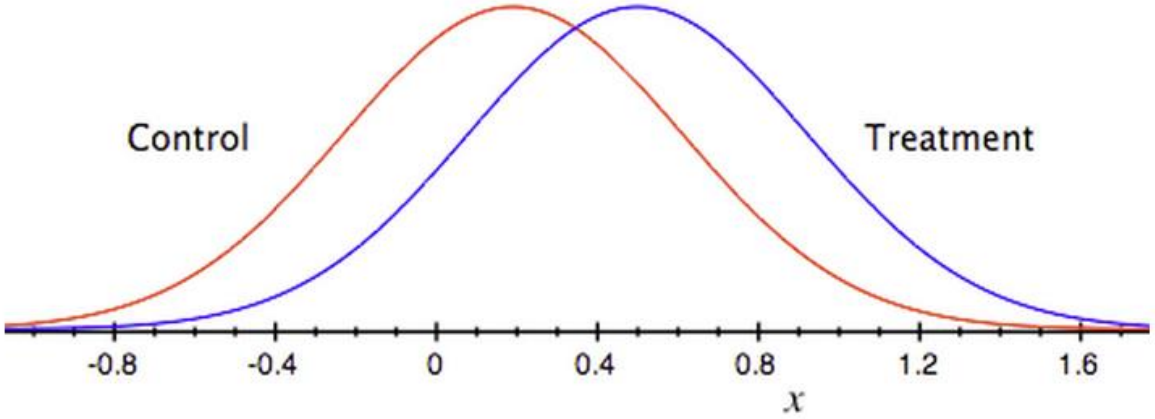
*http://20bits.com/article/statistical-analysis-and-ab-testing



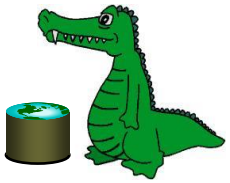
Normality Assumption

The conversion is like a coin flip: heads = "converts" and tails = "doesn't convert"

One can then assume that the sampled conversion rates are normally distributed.



*http://20bits.com/article/statistical-analysis-and-ab-testing



z-scores and One-tailed tests

We define a new r.v. $X = p - p_c$. The null hypothesis becomes

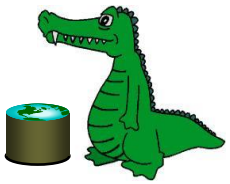
$$H_0: X \leq 0$$

z-score for X: the distance from the population mean to X in terms of standard deviations ($\pm\sigma$)

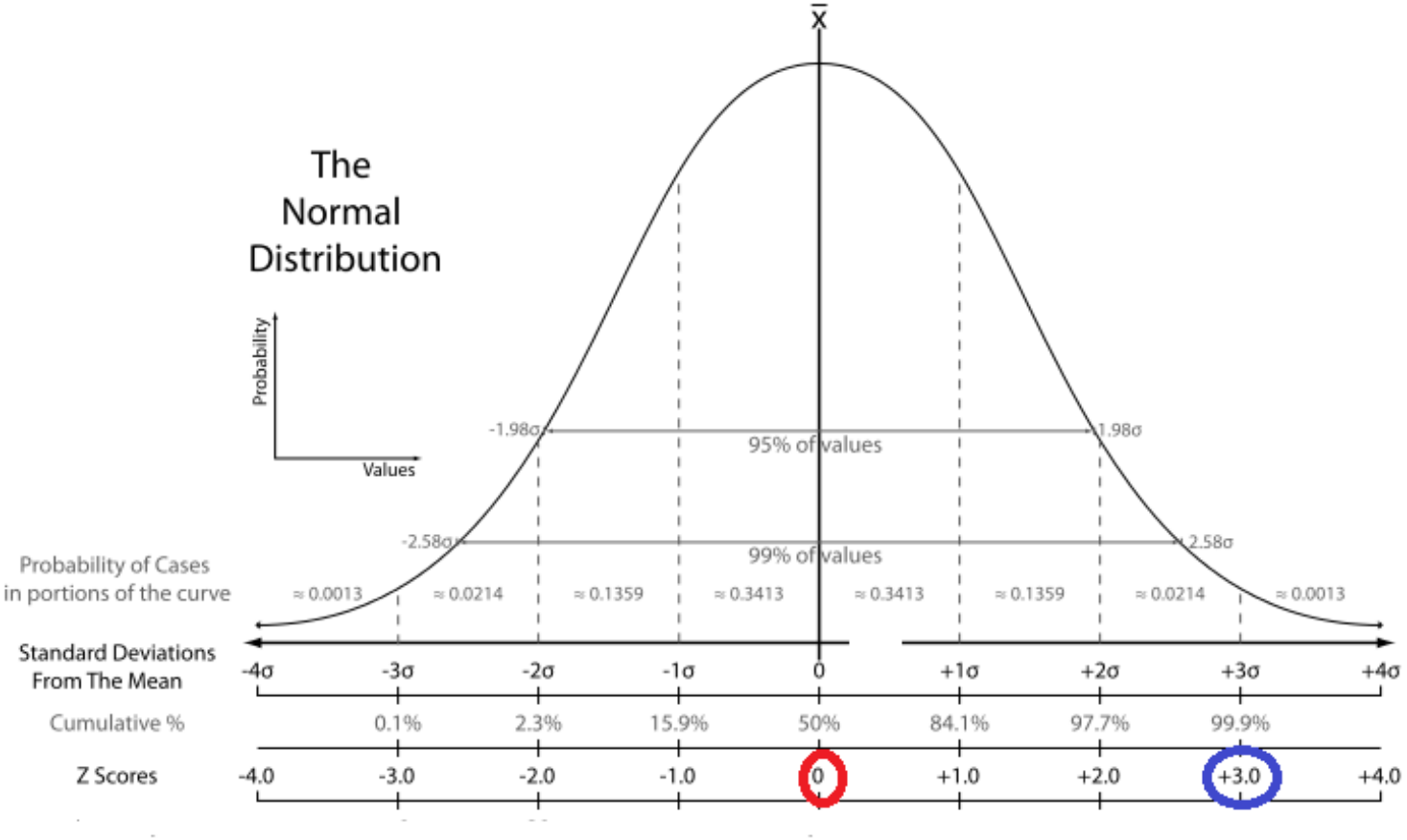
$$z = \frac{\bar{p} - p_c}{\sqrt{\frac{\bar{p}(1-\bar{p})}{N} + \frac{p_c(1-p_c)}{N_c}}}$$

We only care about the positive tail of the normal distribution ($H_a: X > 0$)

*http://20bits.com/article/statistical-analysis-and-ab-testing

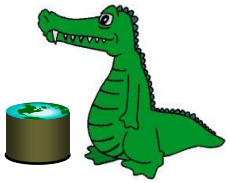


z-scores and One-tailed tests



We reject H_o if the experimental conversion rate is significantly higher (95% confidence i.e. the z-score > 1.65) than the control conversation rate.

* http://www.statisticshowto.com/how-to-calculate-a-z-score/



Summary

Exploratory Data Analysis

- Data types
- Normality assumption and CLT
- Hypothesis Testing