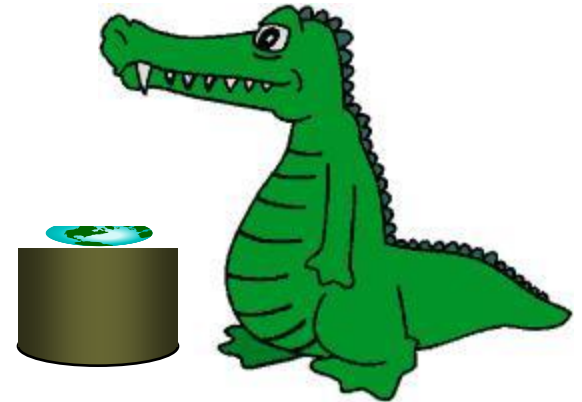


Exploratory Data Analysis





Data – It's numeric*

QUANTITATIVE DATA:



Discrete data:

- There are 3 cones
- Cone 1 has 2 scoops

Continuous data:

- Cone 3 weighs 79.4 grams
- cone 2 ice cream is at 8.3°F

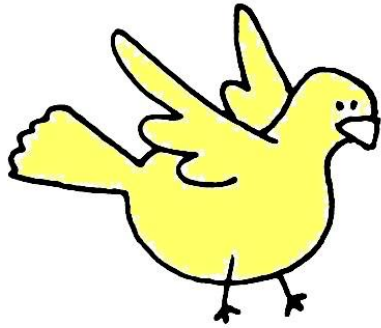
Discrete variables – only a few possible values and no in-between values

Continuous variables – several possible values and in-between values

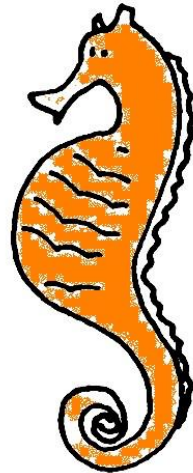


Data – It's descriptive*

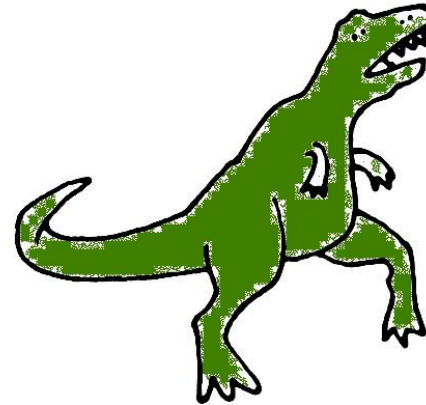
CATEGORICAL DATA:



I am a bird.
I am yellow.
I am awesome.



I am a seahorse.
I am orange.
I am super awesome.



I am a T-rex.
I am green.
I am extinct.



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:

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

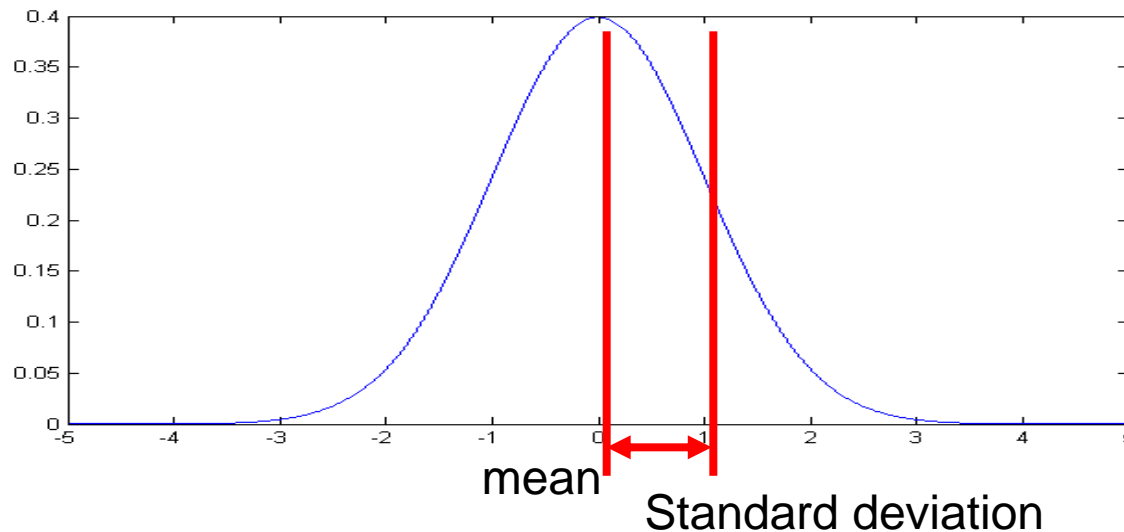
The **standard deviation** is the square root of variance.

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)

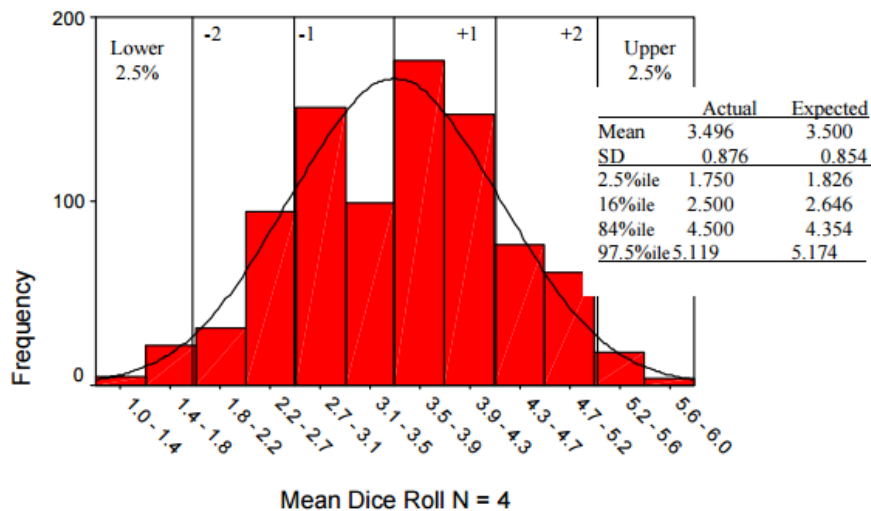




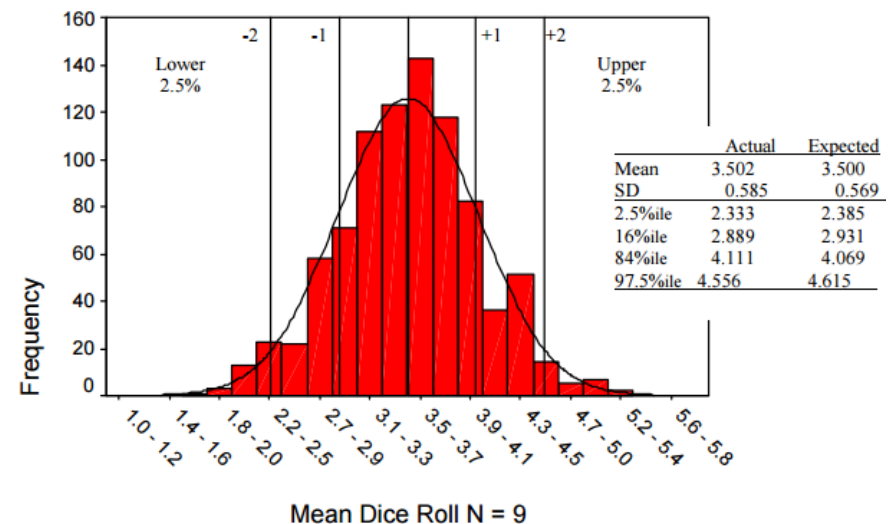
Central Limit Theorem

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

with 900 Replications

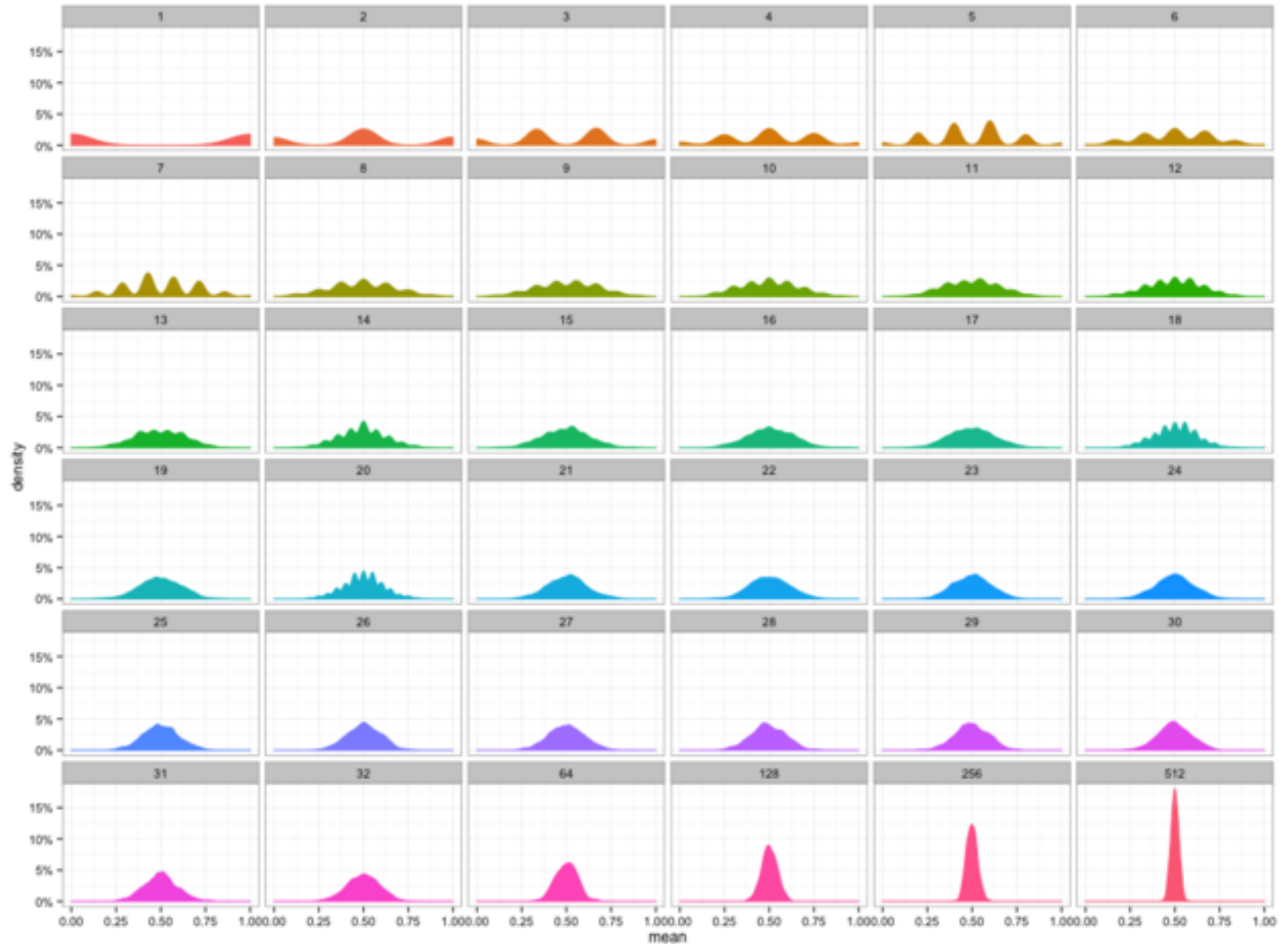


with 900 Replications





Central Limit Theorem





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.

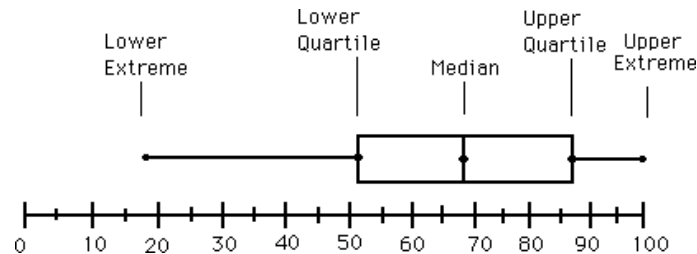
But, they work reasonably well for data that are not normally distributed as long as the samples are not too small (because of CLT).



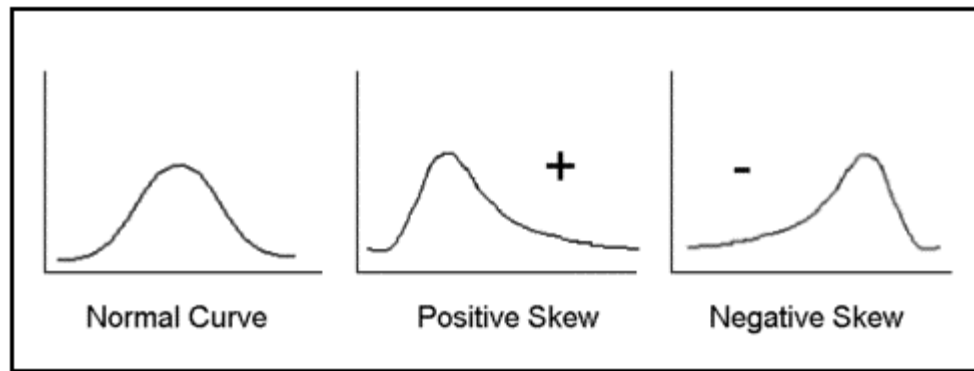
Normally distributed?

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



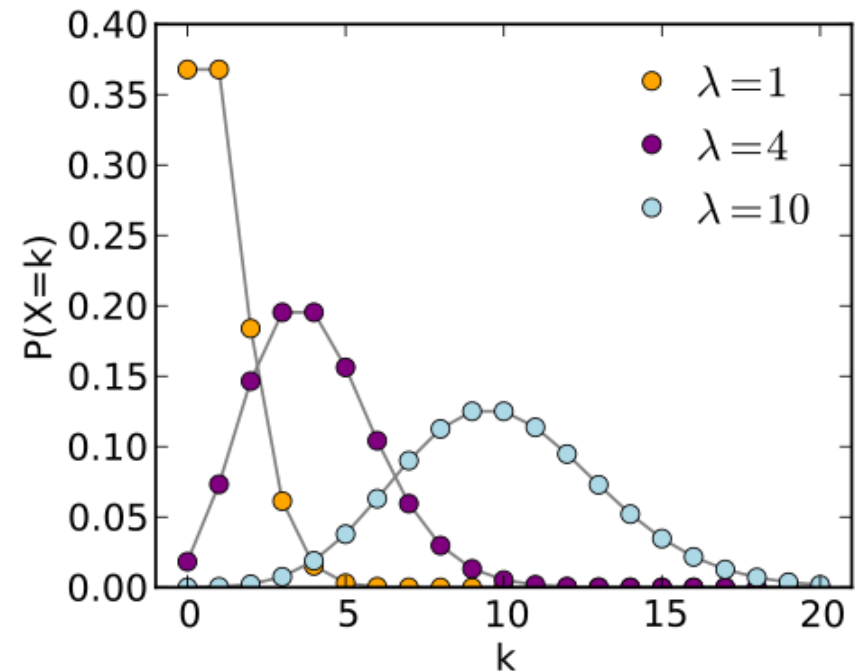
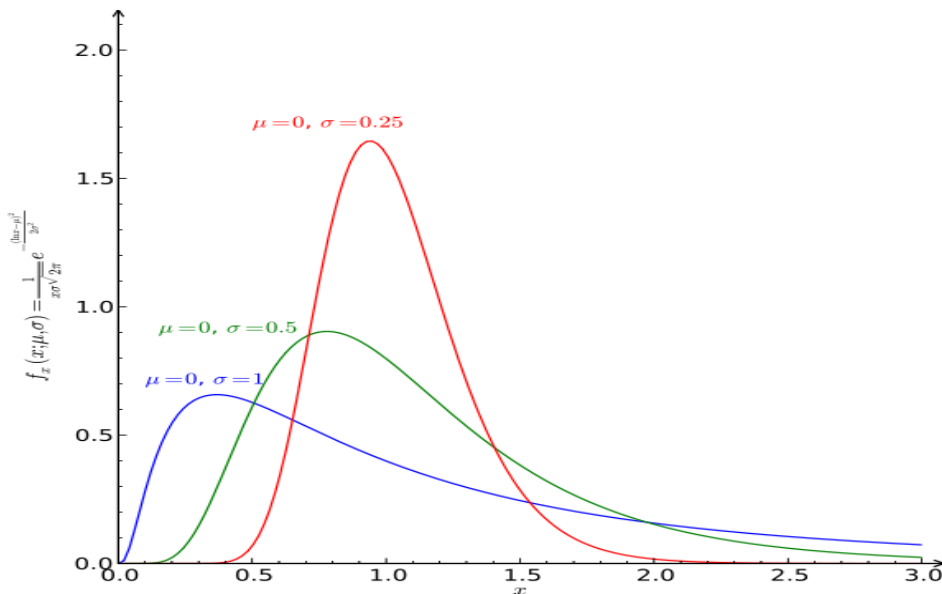
Whenever its asymmetric, the data cannot be normal. The histogram gives even more information





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	Mean	SE Mean	StDev
	Count			
Energy Cost	25	330.6	30.8	154.2



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!



Null & Alternate Hypothesis

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

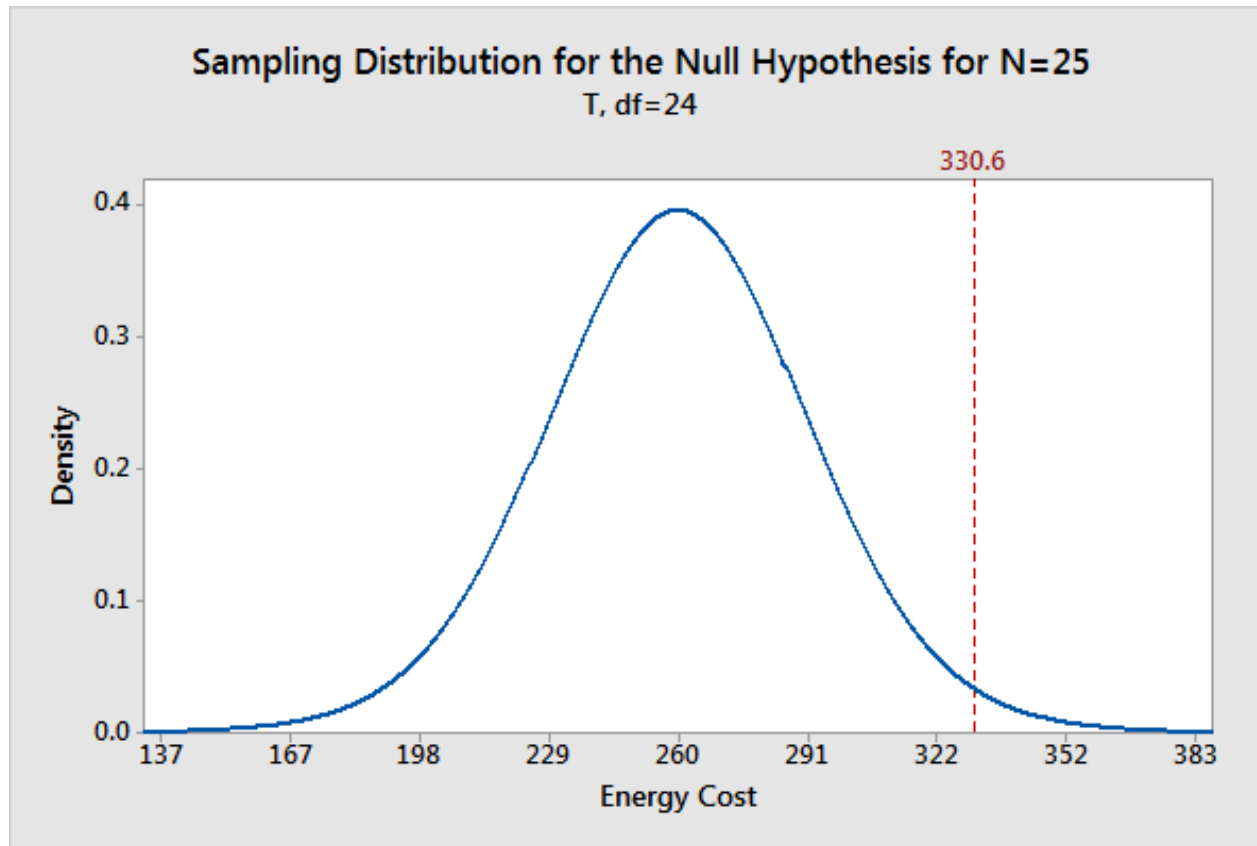
The **alternative hypothesis** (H_a): The population mean (330.6) differs from the hypothesized mean (260).

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

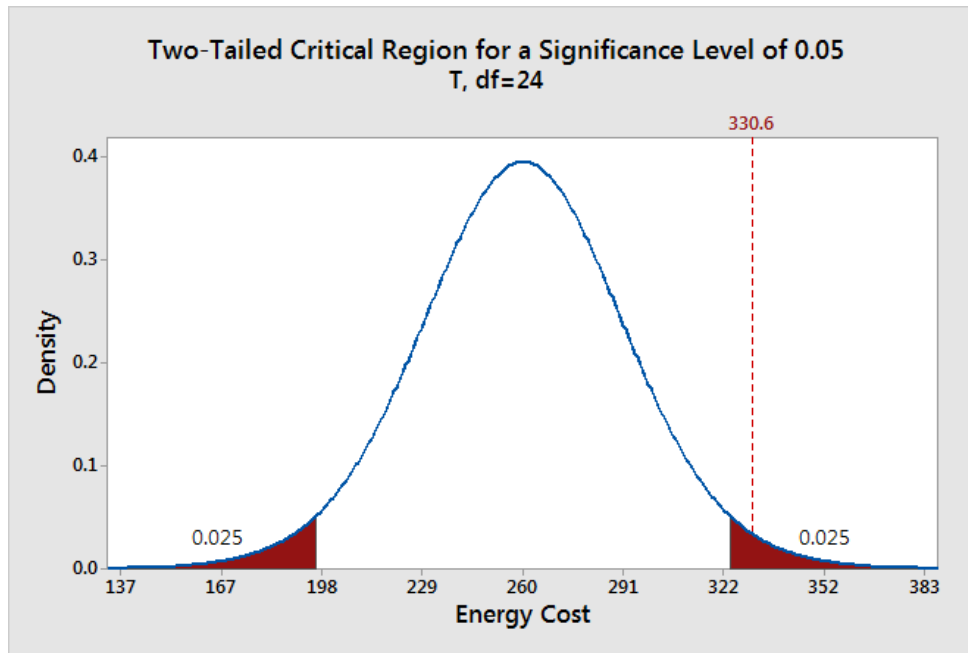
Goal is to determine whether our sample mean (330.6) is **significantly different** from the null hypothesis mean (260)





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

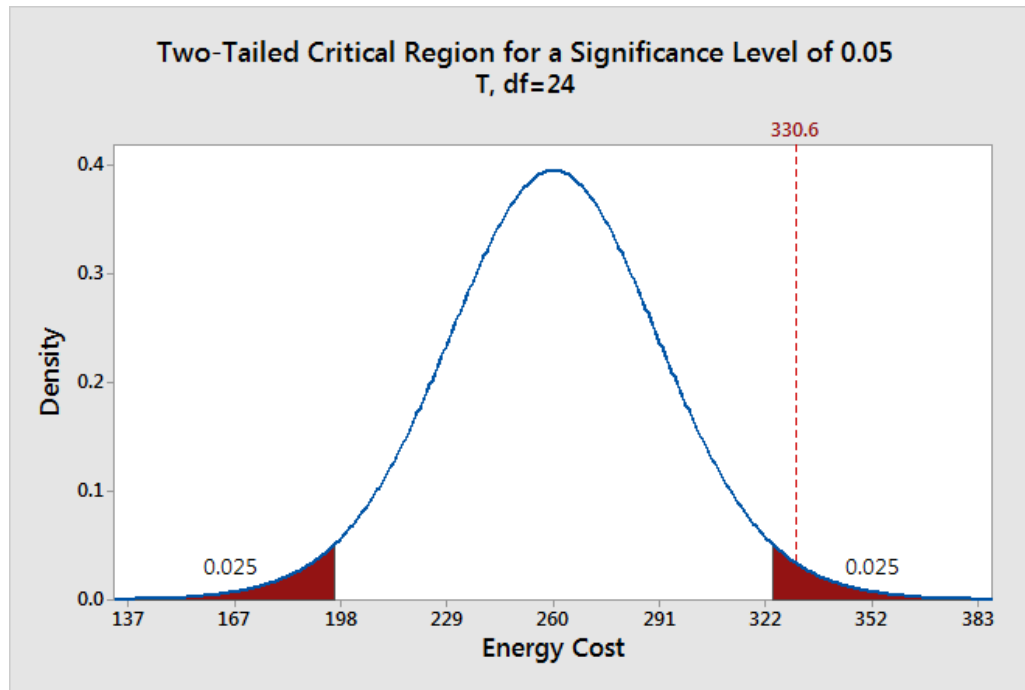


Our sample mean (330.6) falls within the critical region, which indicates it is **statistically significant** at the $\alpha = .05$ level



Significance Level α^*

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

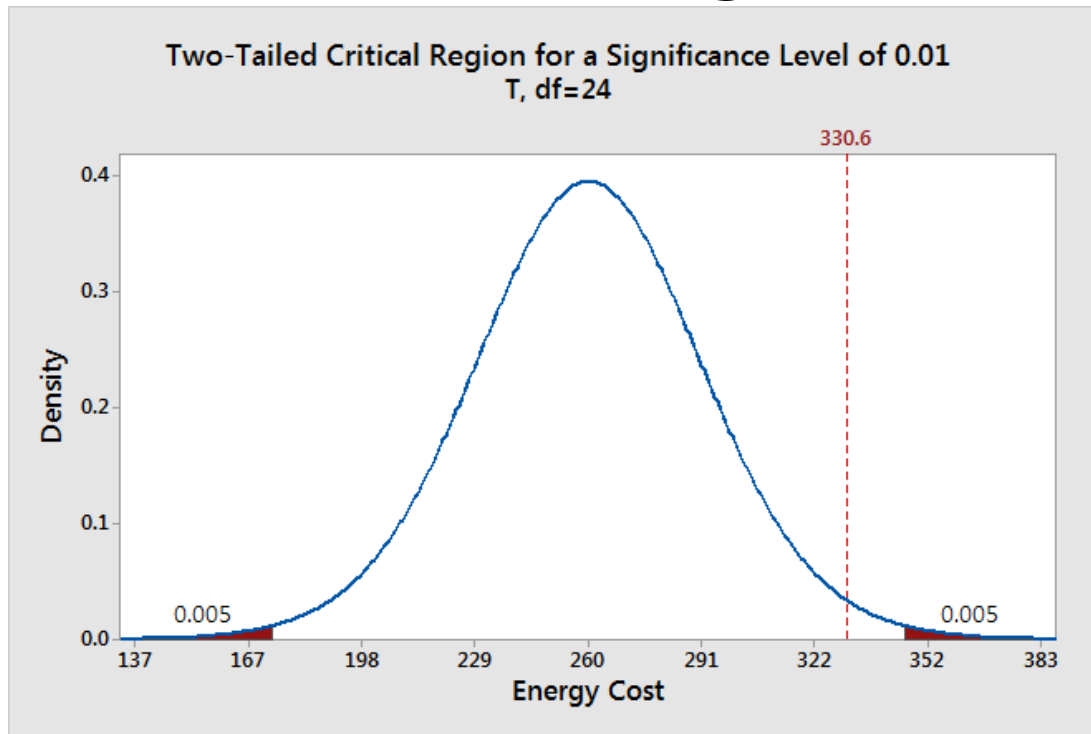


How far out do we draw the line for the critical region?



Significance Level α^*

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



You need to choose the significance level before you begin your study: Common α -values are .01, .05, etc.



What are p -values

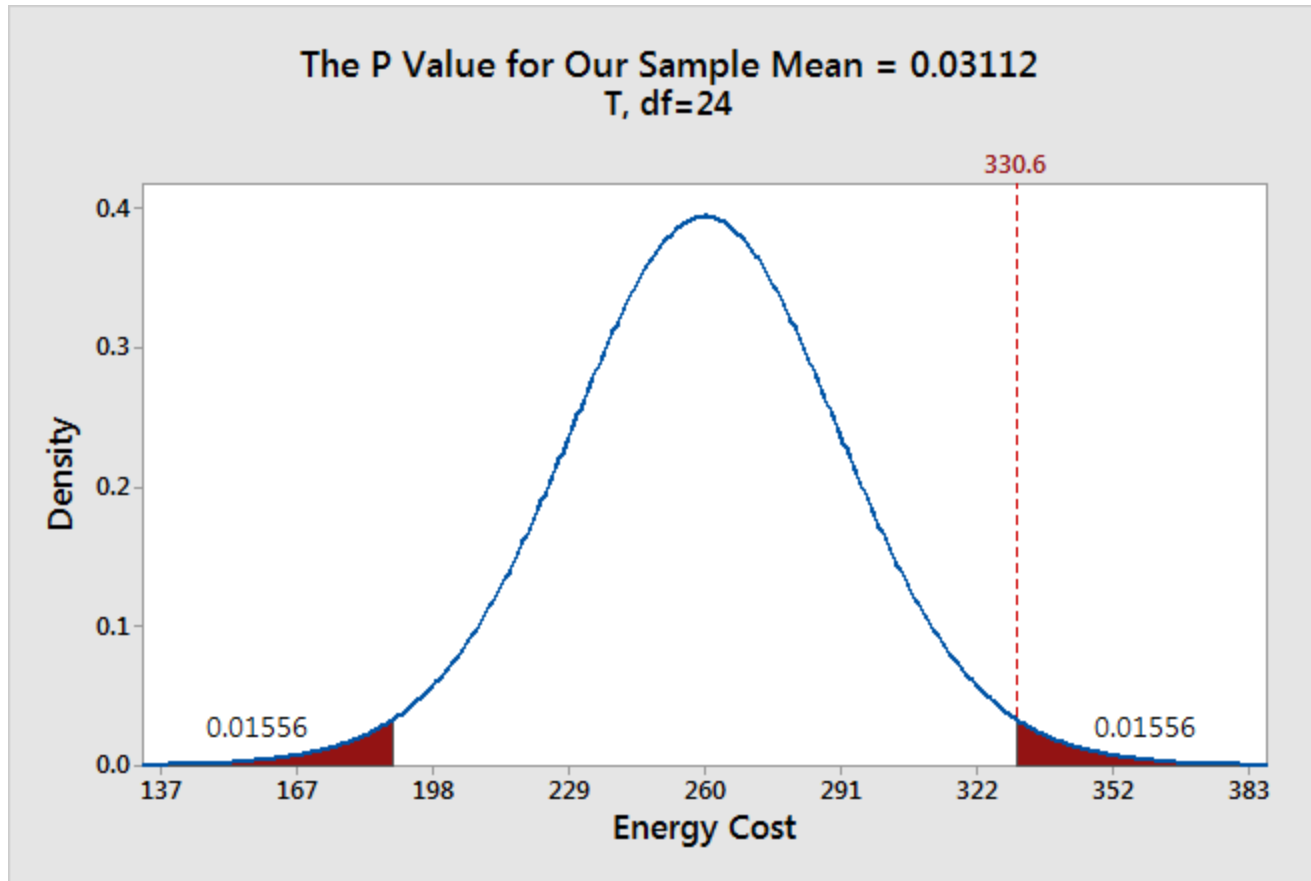
p -value represents the probability of obtaining an effect at least as extreme as the one in your sample data, assuming the truth of the null hypothesis.

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).



What are p -values



If $p\text{-value} \leq \alpha$, we reject the null hypothesis.



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.**