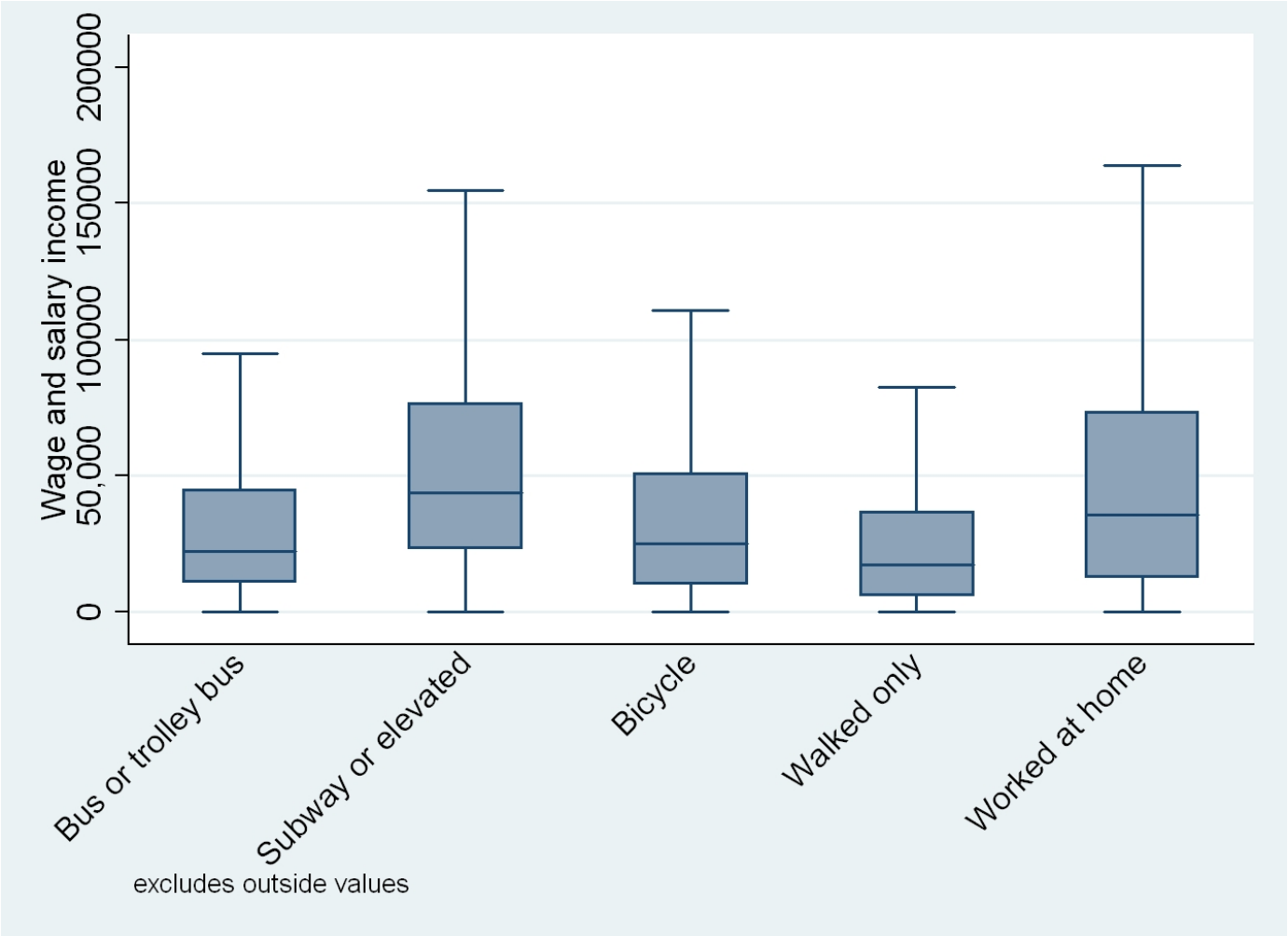
Analysis Toolpack on a Mac

- It seems Excel has done away with the Analysis Toolpack on Macs
- They have worked with another company to provide a close (and free) substitute
- It is called StatPlus:mac LE and can be downloaded from:

  http://www.analystsoft.com/en/products/statplusmacle/

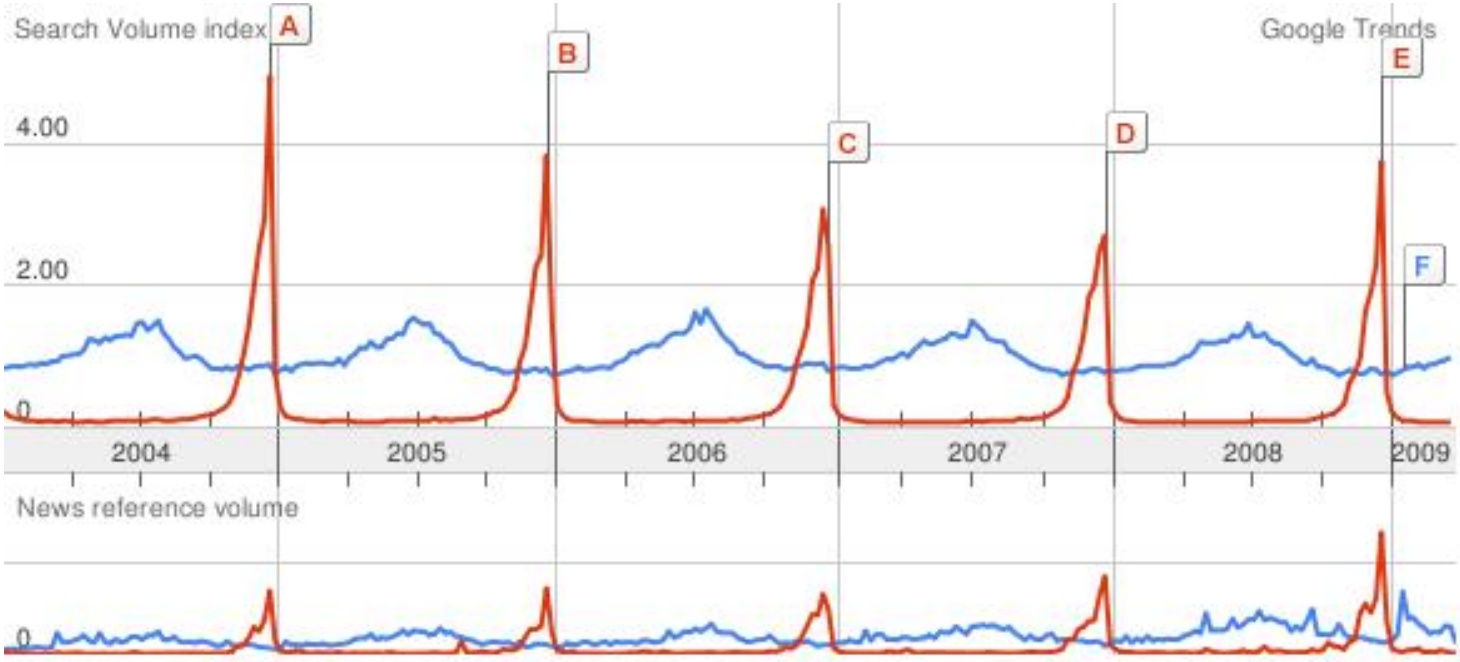
- It is designed to match up quite closely with the PC analysis toolpack

Summary Statistics as a Graph: The Box Plot



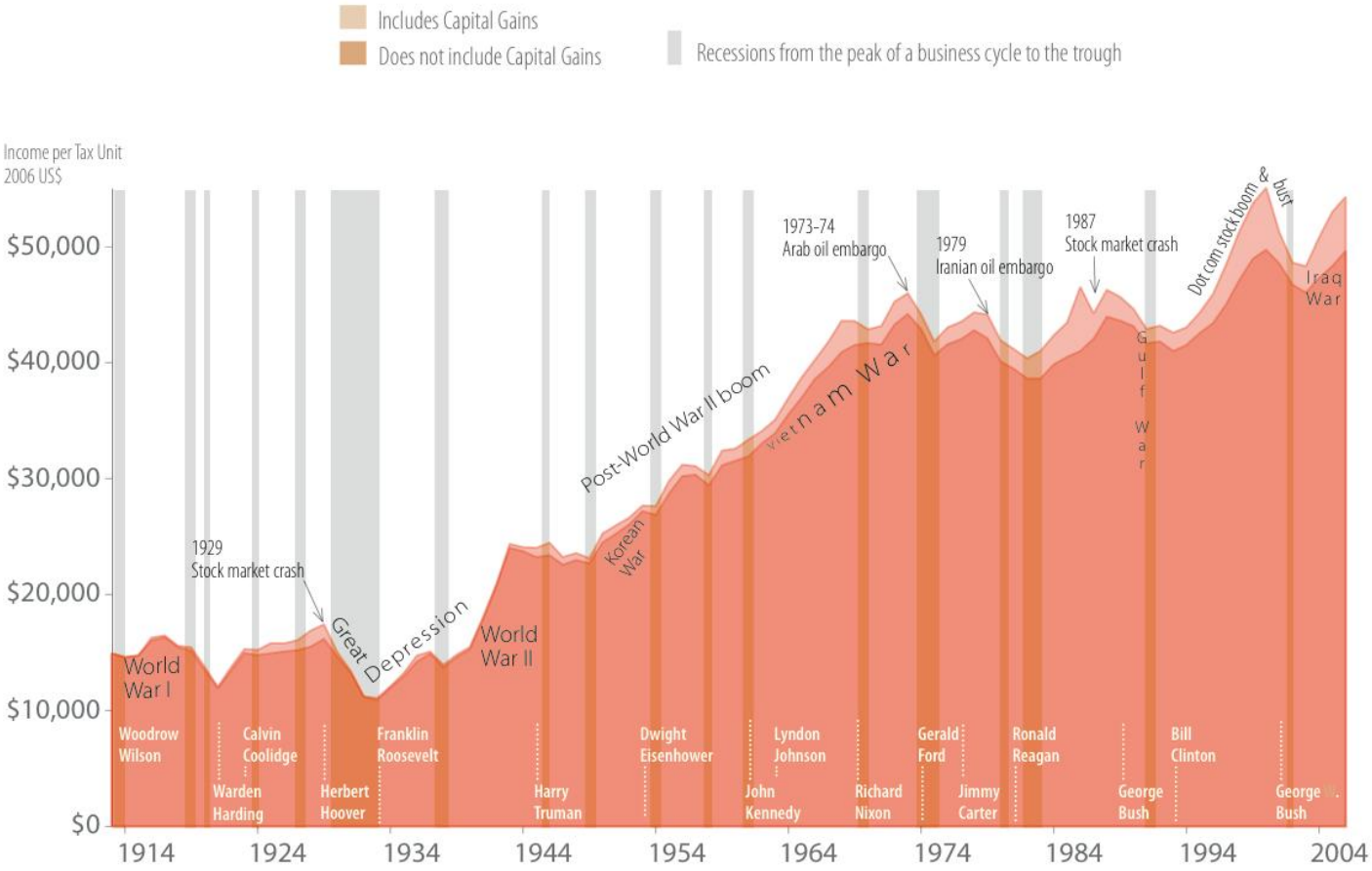
Box plot of income by form of transportation used, 2008 American Community Survey

Some Other Examples of Visual Representations of Data



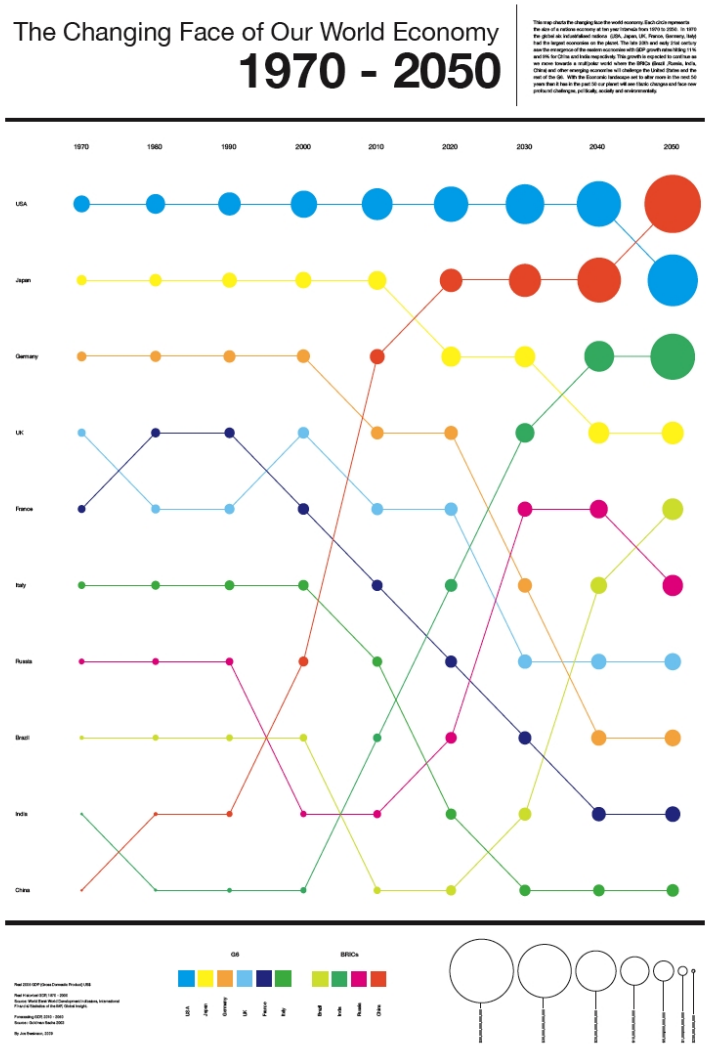
Google Trends data for the phrase "ice cream" (blue line) and the word "Santa" (red line).

Some Other Examples of Visual Representations of Data



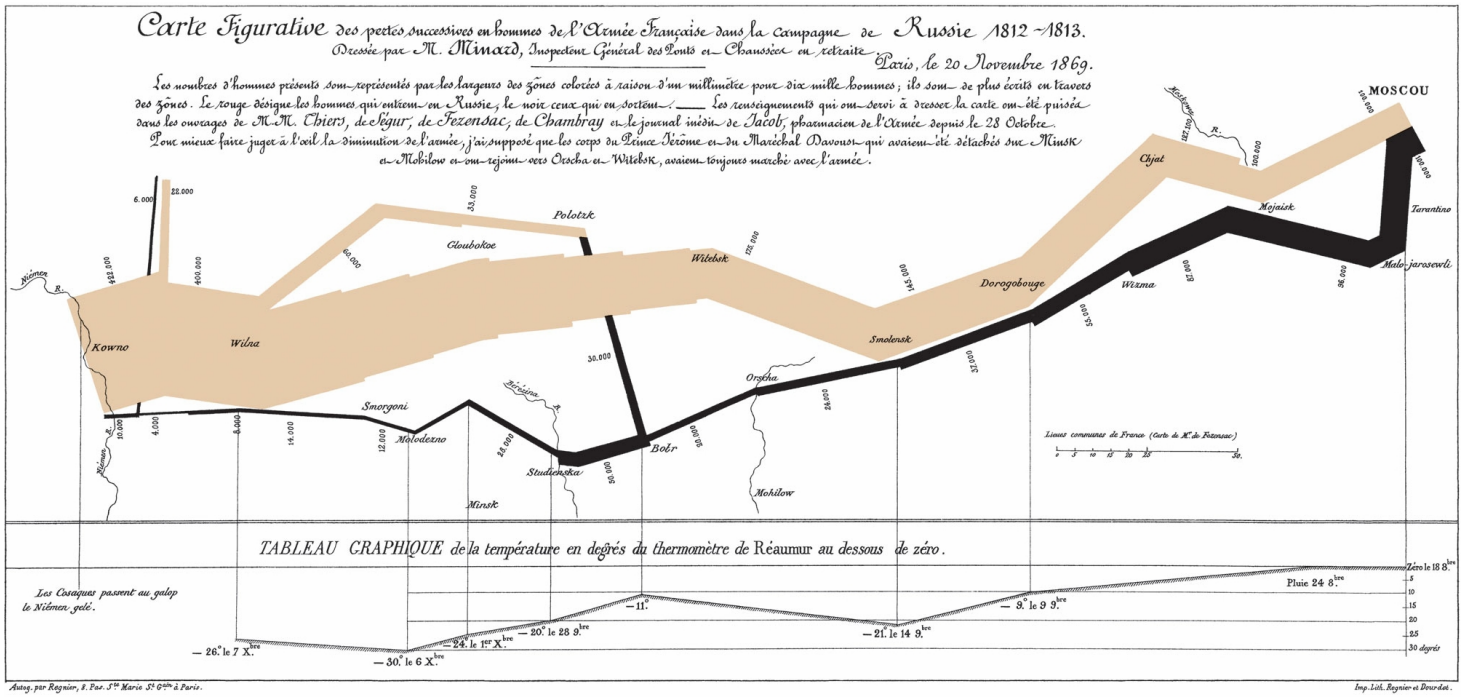
From visualizingeconomics.com

Some Other Examples of Visual Representations of Data



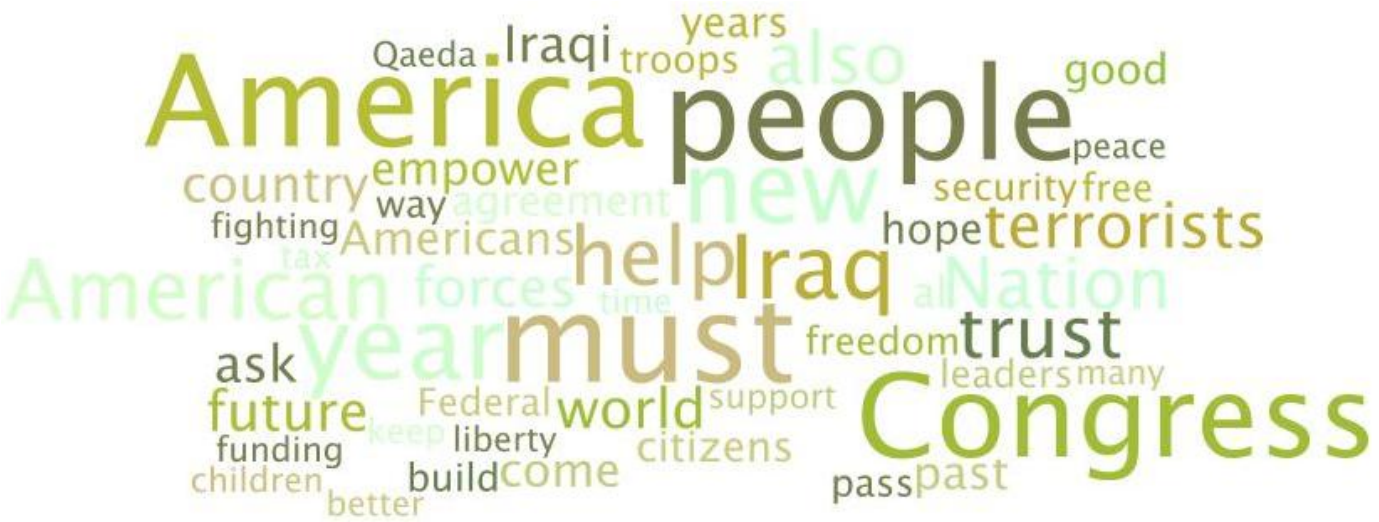
From joeswainson.blogspot.com

Some Other Examples of Visual Representations of Data



Map of Napoleon's Russian campaign of 1812, Charles Joseph Minard (1861)

Some Other Examples of Visual Representations of Data

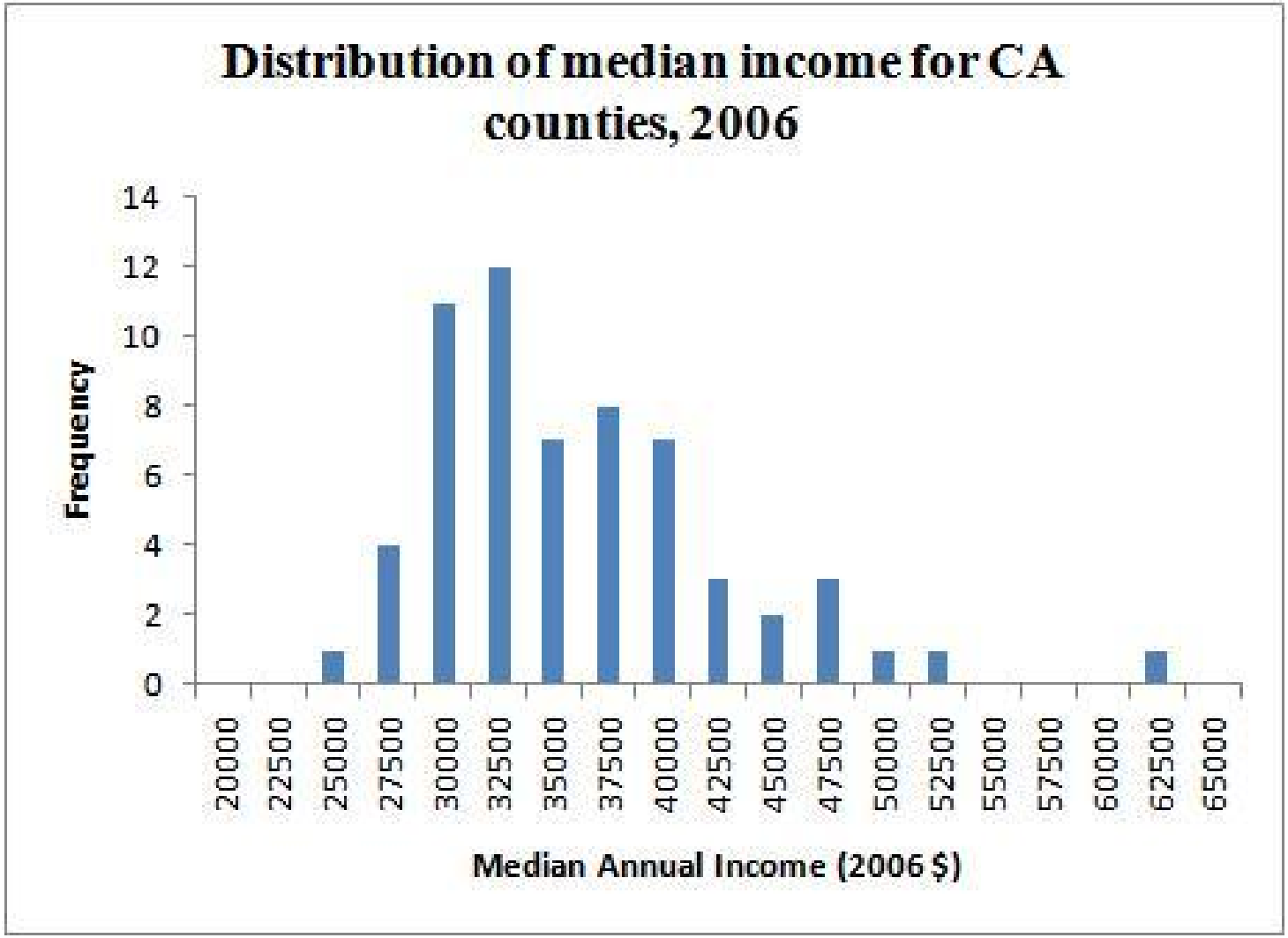


Wordle generated from Bush's 2002 State of the Union address (after 9/11).

Some Other Examples of Visual Representations of Data

Wordle generated from Obama's 2009 State of the Union address (after start of recession).

Review of Univariate Summary Statistics



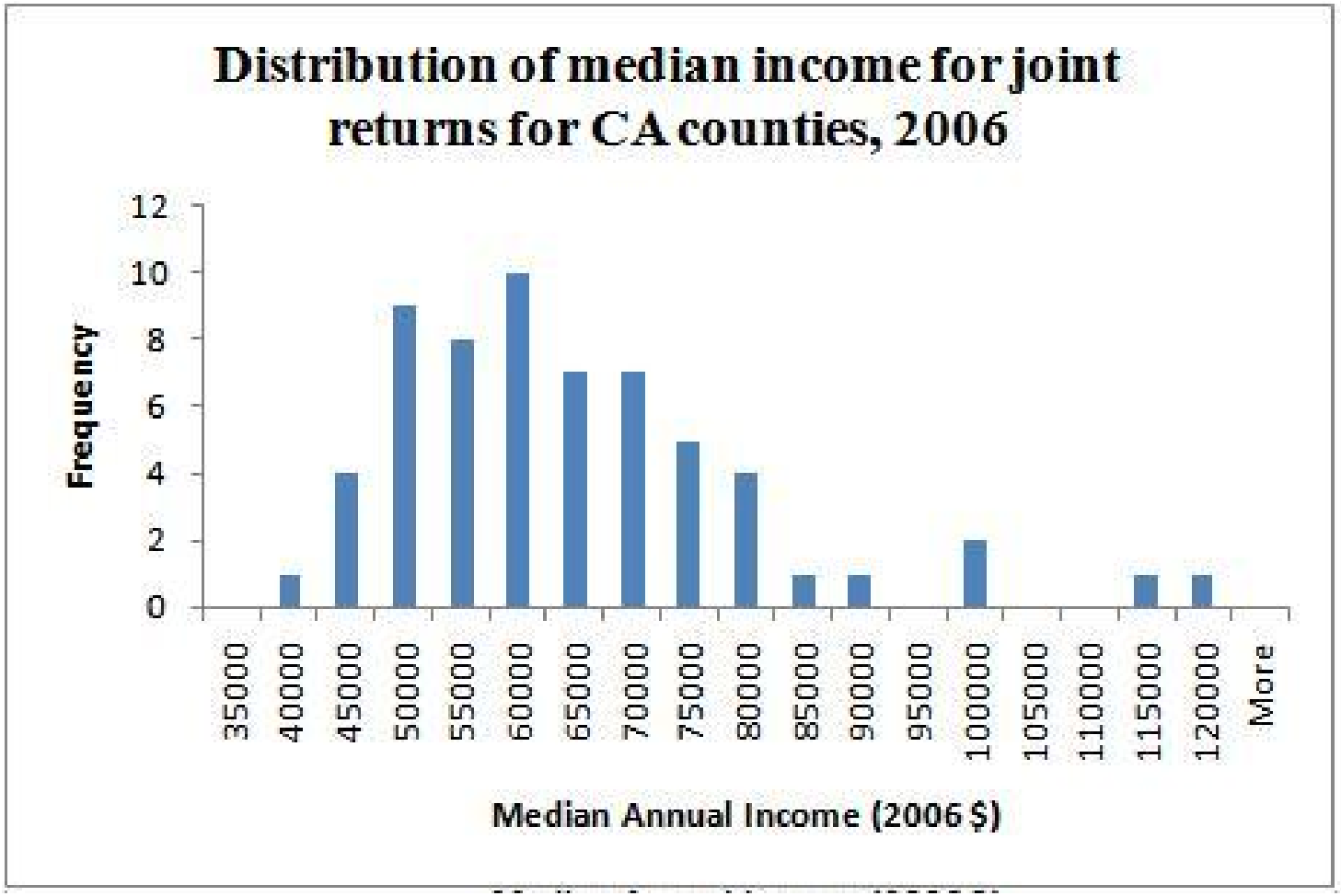
Review of Univariate Summary Statistics

Median Income (all returns)	
Mean	34831.13115
Standard Error	908.0839061
Median	33103
Mode	28417
Standard Deviation	7092.362034
Sample Variance	50301599.22
Kurtosis	2.67267105
Skewness	1.338008719
Range	38652
Minimum	23557
Maximum	62209
Sum	2124699
Count	61
Confidence Level(95.0%)	1816.438244

Review of Univariate Summary Statistics

Median Income (joint returns)	
Mean	62308.5082
Standard Error	2042.739224
Median	58959
Mode	#N/A
Standard Deviation	15954.30336
Sample Variance	254539795.8
Kurtosis	2.15419599
Skewness	1.284590168
Range	79044
Minimum	37582
Maximum	116626
Sum	3800819
Count	61
Confidence Level(95.0%)	4086.086785

Review of Univariate Summary Statistics



Univariate Statistical Inference

Univariate Statistical Inference

- Statistical inference: using sample statistics to make inferences about the population
- For univariate data, this means using the sample average to make inferences about the population mean
- Examples of why we do this: polls to infer public opinion, water samples to assess water quality, etc.

Steps for Statistical Inference

The basic approach to making an inference about the population mean is the following:

1. Form a hypothesis about the population mean
2. Create a test statistic
3. Use the test statistic to decide whether to reject the hypothesis
4. Interpret the result

Some Definitions

- **Random variable**: a variable that can take on a variety of values, each with some particular probability, we'll denote a random variable with X
- **Realization of a random variable**: an observed outcome for a random variable, for example the outcome of a coin flip turning out to be heads, we'll donate a realization of a random variable with x
- **Population**: the set of all realizations of a random variable X
- **Sample**: a subset of realizations of X selected from the population $(x_1, x_2, ..., x_n)$

More Definitions

- **Random sample**: a sample where each observations is an independent draw from the same population
- **Independent draws**: the probability of a draw taking on any particular value is not affected by the outcomes of the other draws
- **Population mean**: the average of all possible values of X (which is the expected value of X) in the population, written as either μ or $E(X)$
- **Sample mean**: the average of the n different values of x in a particular sample $(x_1, x_2, ..., x_n)$, written as $\bar{x}$
- Note that the sample mean $\bar{x}$ is a random variable, it will have different values for different samples

The Basic Idea

We want to use a sample to infer whatever we can about the distribution of random variable X at the population level

- What would we like to know about the population?
  - the population mean, μ
  - the population variance, σ^2
  - the shape of the distribution of X, *pdf* (probability density function)
- What information do we actually get to observe?
  - the mean of the sample, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - the standard deviation of the sample, $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - the same statistics for any additional samples we take

The Basic Idea, continued

The basic idea of hypothesis testing is the following:

- Formulate a hypothesis that μ is equal to some particular value, say 100
- If the sample mean is very close to 100, then we won't reject this hypothesis
- If the sample mean is very far from 100, then we will reject the hypothesis
- The tricky part is how to define 'very close' and 'very far'

The Distribution of the Sample Mean

- Remember that the sample mean $\bar{x}$ is actually a realization of a random variable $\bar{X}$
- We'll use the properties of the distribution of $\bar{X}$ to define 'very close' and 'very far'
- It turns out that the sample mean is distributed normally with a mean equal to the population mean of X and a variance equal to the population variance divided by the sample size

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

- This is true even if X isn't normally distributed

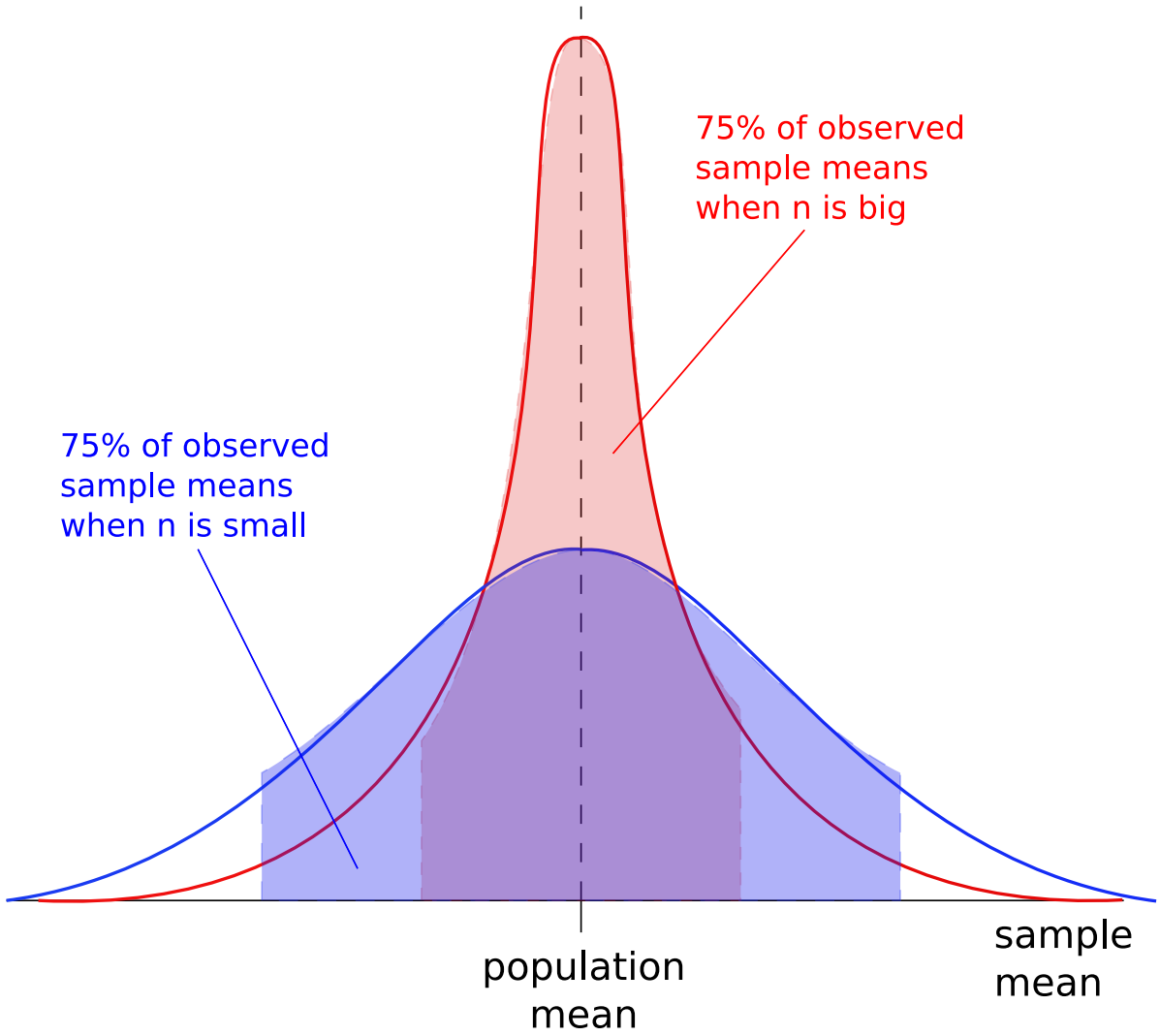
The Distribution of the Sample Mean

- To get a better sense of the distribution of the sample, we'll go through a very simple example
- Let's think about coin flips, we'll call heads '1' and tails '0'
- The set of all possible values is just $(0, 1)$ each with a probability of $\frac{1}{2}$
- The population mean, or expected value of a coin flip, should just be $\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 1 = \frac{1}{2}$
- If we take a sample by flipping a coin a few times, what are we likely to see as the sample mean?
- See distribution-of-sample-mean.xlsx

The Distribution of the Sample Mean

- So the average value of the sample mean should tell us the population mean suggesting that we can use $\bar{X}$ to get an estimate of μ
- For a single sample, it is unlikely that the observed $\bar{x}$ is exactly equal to μ
- The standard deviation of the sample mean, often called the standard error of the sample mean, helps us understand how likely it is that a sample mean will be close to to the population mean
- The smaller the standard error, the narrower the distribution of the sample mean and the better our sample mean is as estimator of the population mean

The Distribution of the Sample Mean



Sample Mean as an Estimator of μ

- $\bar{X}$ is an **unbiased** estimator of μ:

$$E(\bar{X}) = \mu$$

- $\bar{X}$ is a **consistent** estimator of μ:

$$\lim_{n \to \infty} \bar{X}_n = \mu$$

- In some cases, $\bar{X}$ has the **minimum variance** among consistent estimators

Restating the Main Idea

Now we can state our hypothesis testing procedure a little more formally:

- Form a hypothesis that μ is equal to a particular value μ_0
- Calculate sample mean and sample standard deviation
- Given the sample standard deviation, what would the probability be of observing $\bar{x}$ if the true population mean is μ_0?
- If the probability is high, don't reject the hypothesis
- If the probability is very low, reject the hypothesis

Test Statistics

- We get these probabilities by constructing a standardized test statistic
- If we knew the true population variance σ^2, we would calculate a z-score:

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

- Since we don't know σ, we have to use the sample standard deviation s and calculate a t-score:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \sim T_{n-1}$$

What a Test Statistic Tells You