Fun with Statistics:

You'll Probably Have a Good Time

Zoology 306L

TA: Nate Wehr

Statistics

- What is a Statistic?
- What kinds have we used so far?
- What are their purpose?
- What are their limitations?

Inference Statistics

- Reason for inference statisics
- Limitations and assumptions
- Pitfalls
  - Overfocused on the analysis

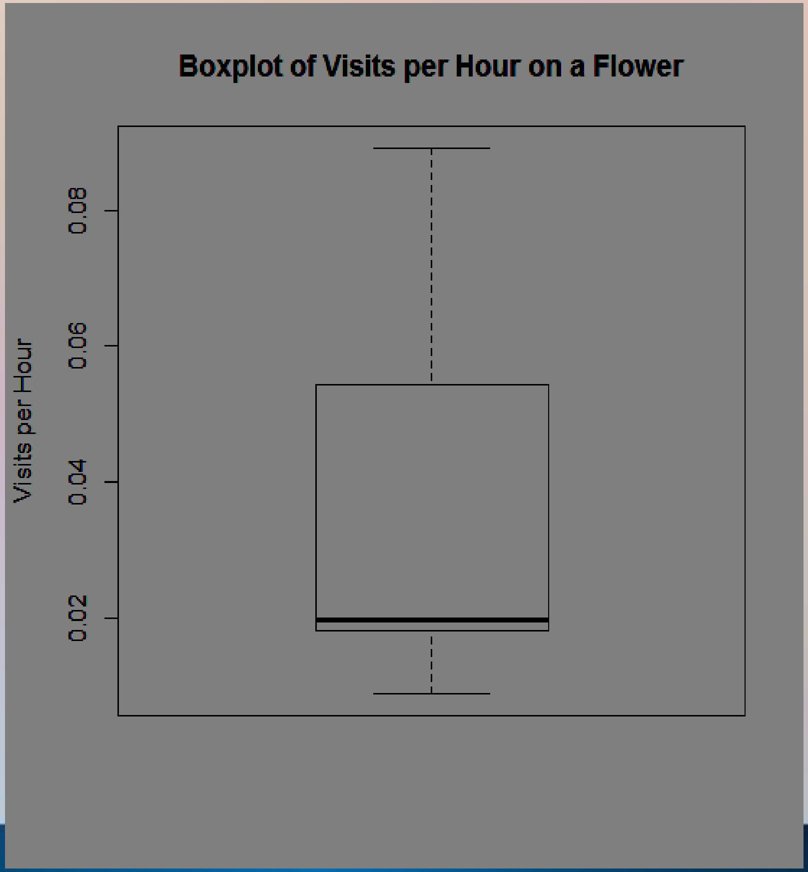
Moving from Descriptive to Inference

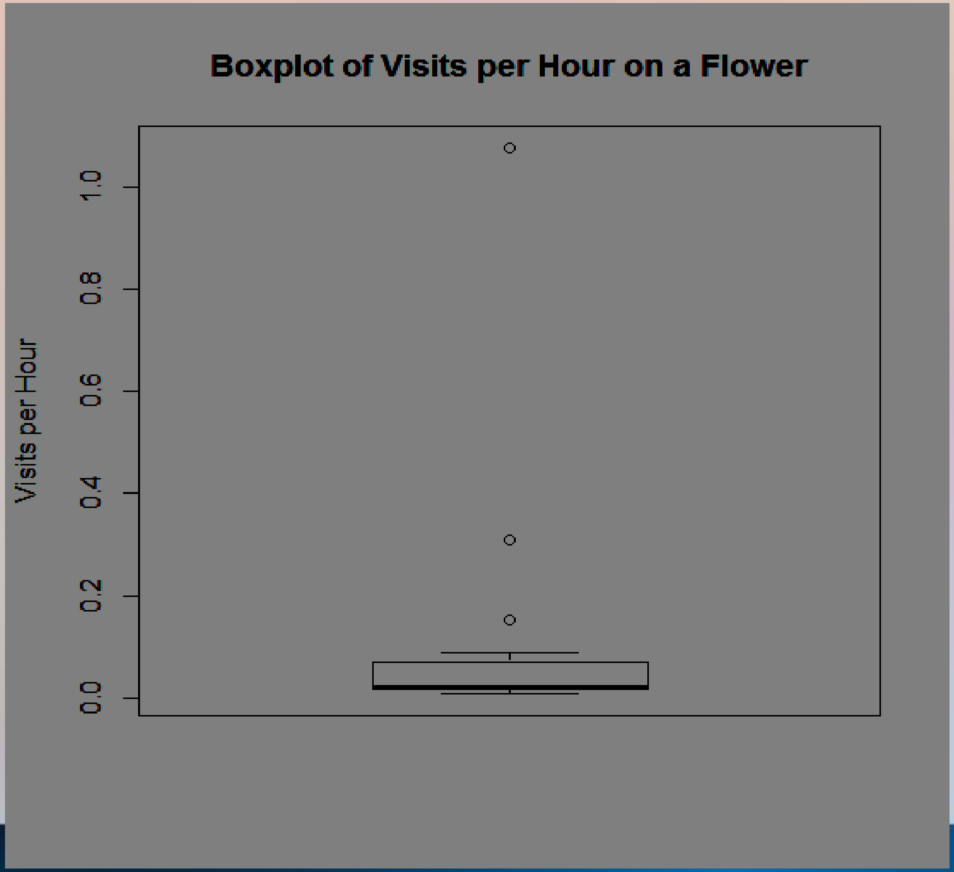
- Mean vs. Median
- Symmetry vs. Skewed
- Outliers
- 5-number summary

Boxplots!

- Min 1st Quantile Median 3rd Quantile Max
- 0.0090 0.0181 0.0198 0.0543 0.0891

1.5 x IQR rule



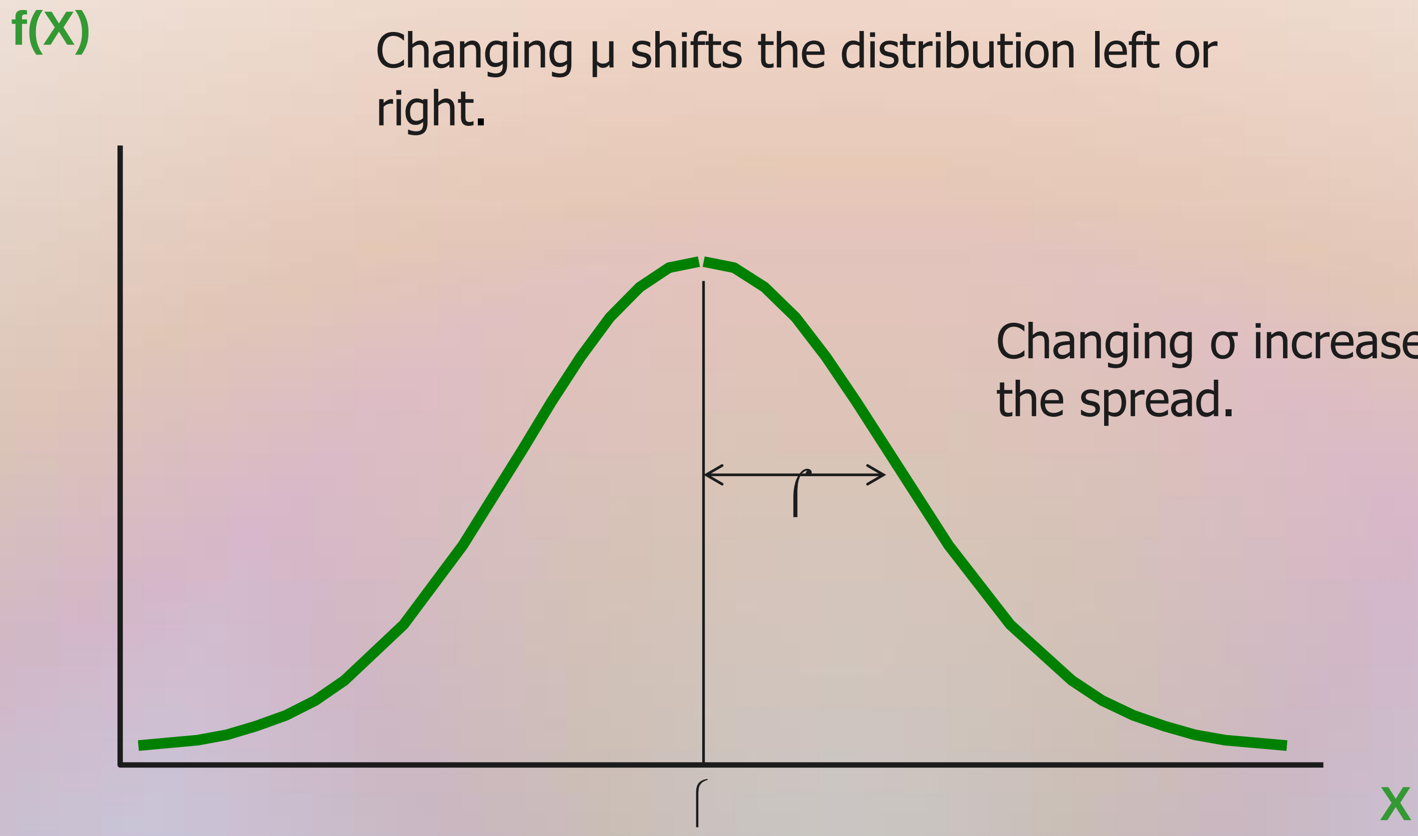


Moving from Descriptive to Inference

- Variance (σ^2)
- Standard deviation (σ)
- Normality
- 68-95-99.7 rule

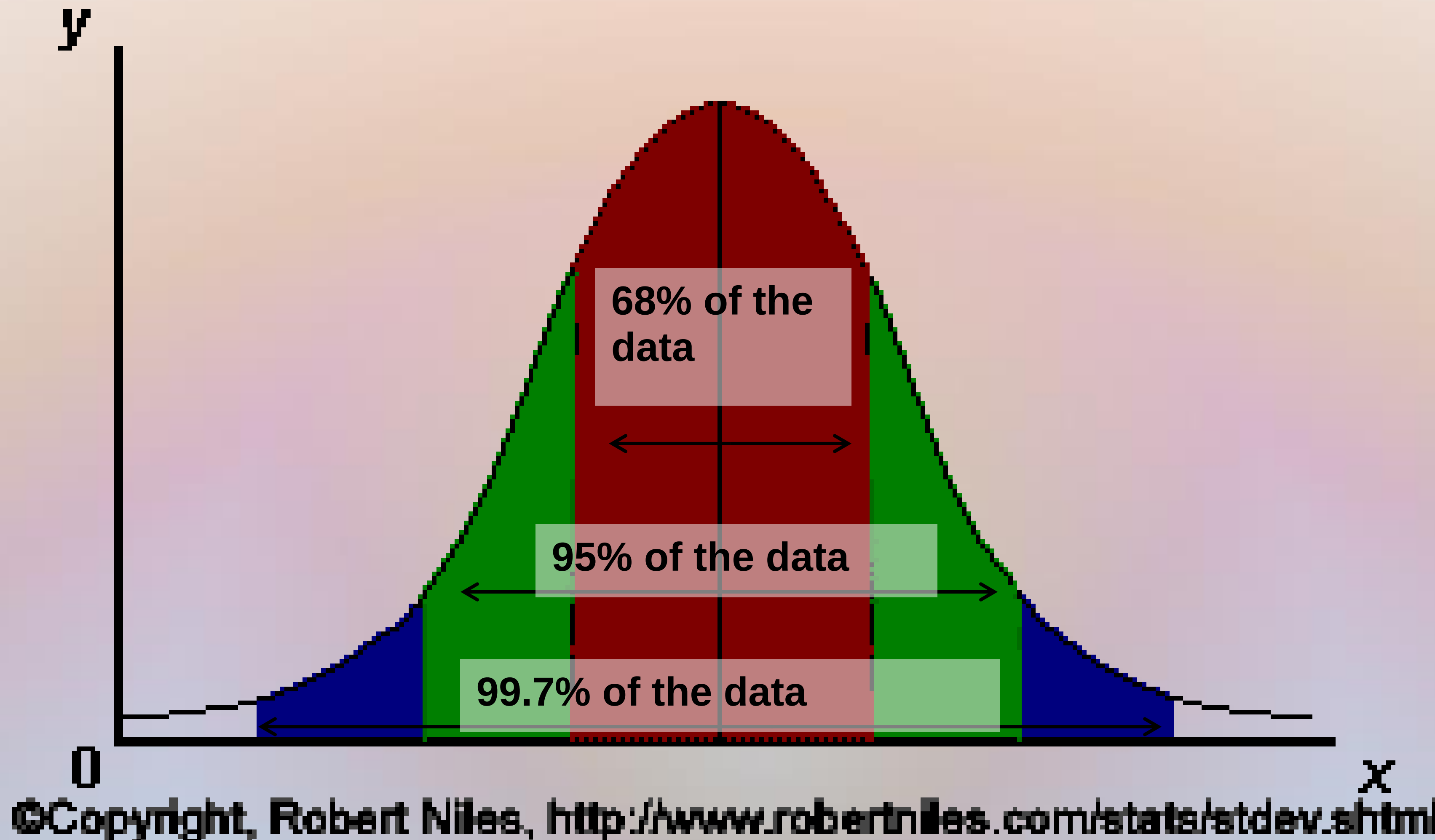
Normal Distribution

Changing μ shifts the distribution left or right.

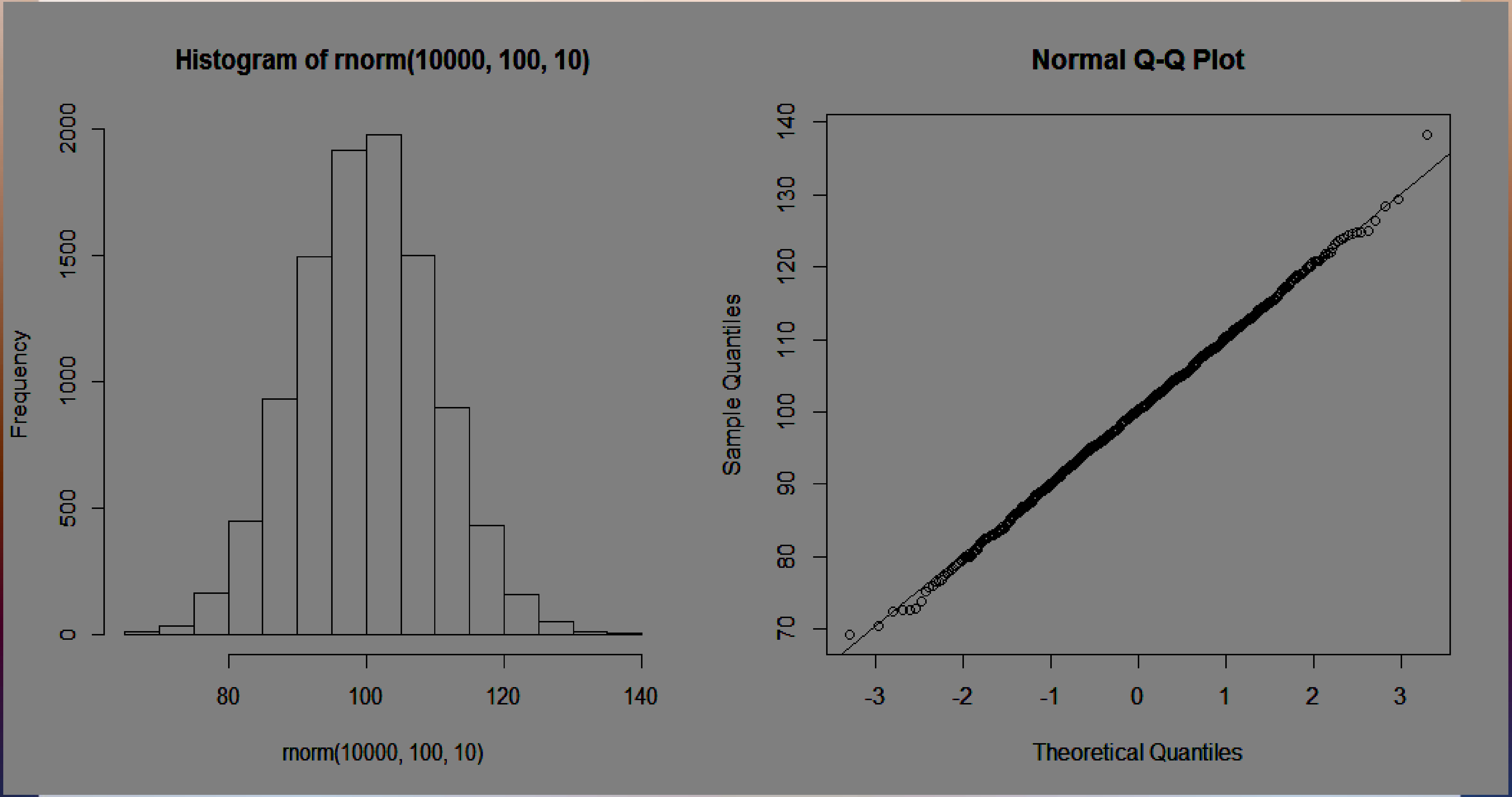


Changing σ increases or decreases the spread.

68-95-99.7 rule



Assessing Normality



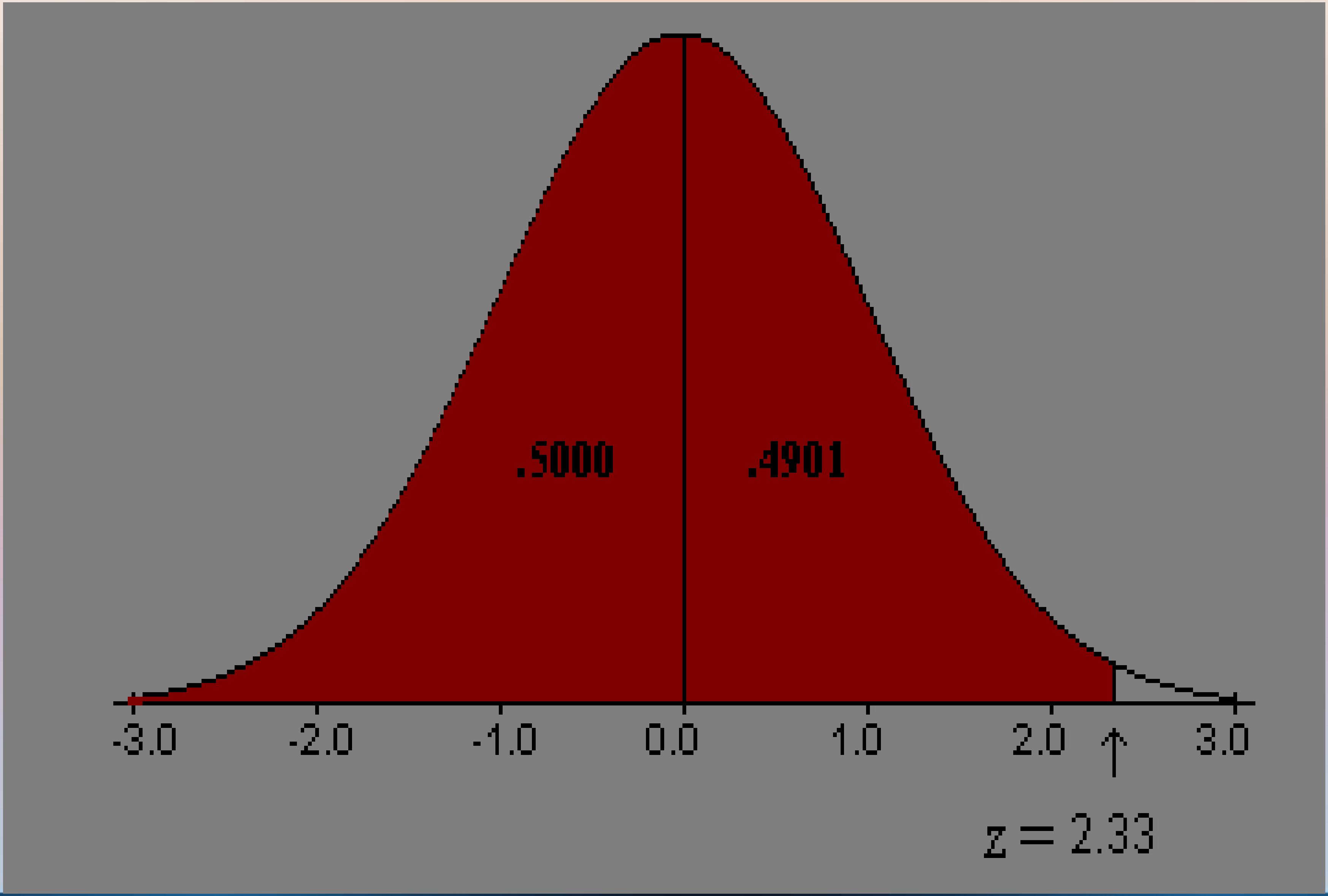
Inference Testing

- Variation in samples
  - How do we know it is a real difference?

Statistical Hypothesis Testing

- Experimental/Study Hypothesis vs. Statistical Hypothesis
- Null Hypothesis (i.e. H_0: $x = \mu$)
- Alternative (i.e. H_a: $x \neq \mu$)
- Standard Error: (SE_x) from estimate of σ
  - Comparing your mean to the Population Mean

Probability



Critical Values

- Point at which you reject the null hypothesis
  - Each test will have one

- Confidence Level (α)
  - Why $\alpha = 0.05$?

Errors with Probabilities

- Either correct or not
- Type I error
  - False positive
- Type II error
  - False negative

- Which is worse?

Types of Tests used in this class

- Chi-Squared
    - Used often for choice experiments
    - Data are compared to an expected value
    - Data are categorical and discrete (whole numbers)

- T-test
    - Used to compare means of two separate groups
    - Data are categorical and continuous

- Paired T-Test
    - Used to compare means of the same group twice
    - Used to determine the affect of a treatment on one group
    - Data are categorical and continuous

- ANOVA
    - Used to compare the means of three or more separate groups
    - Data are categorical and continuous

- Simple Linear Regression
    - Used to find correlations between two numerical, continuous variables

Chi-Squared (χ^2)

- Used for Frequency Data
  - Best when categorical variables are important

- Null hypothesis is no association (or difference) between variables

Chi-Squared (χ^2)

- Assumptions:
  - Random Sampling
  - Independant Observations (can only exist in one category)
  - Large Sample size → Expected frequencies need to be large

Chi-Squared (χ^2)

- $\chi^2 = \sum[(\text{Observed-Expected})^2/\text{Expected}]$
- One-Way: df = n-1
- Two-Way: df = (row-1)(column-1)= (r-1)(c-1)
  - (If one of these is success= t test)
- Use Table or Software for p-value

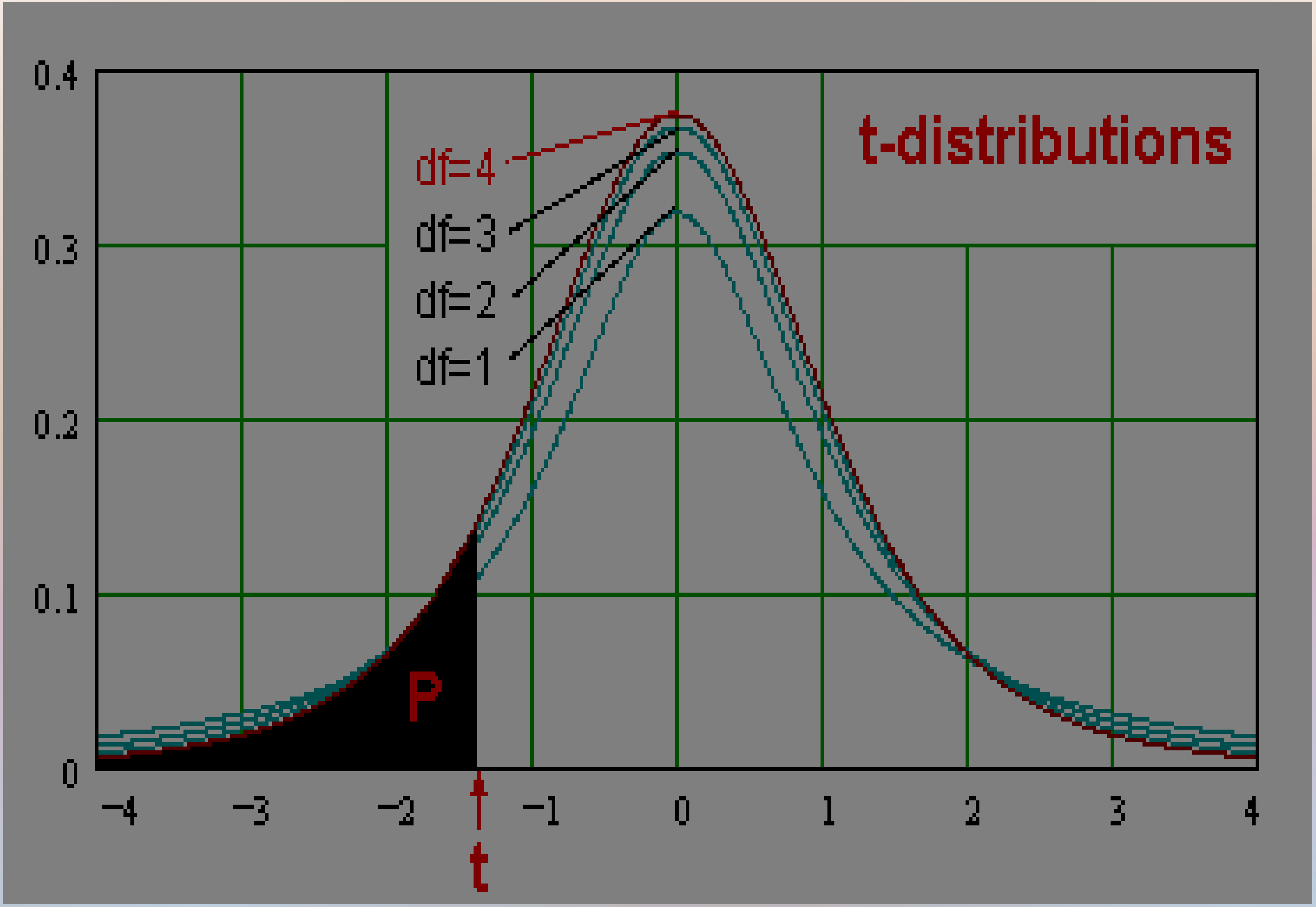
t-test (One Sample)

- Used to compare to population mean.
  - Simple random sample

- t-distribution comes from sample with unknown std deviation

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

t distribution



t-test (One Sample)

$$t = \frac{\bar{x} - u_0}{s/\sqrt{n}}$$

- df = n-1

- Use a table to look up p-value or software

t-test

- Assumptions of t-tests
  - Independant Observations
  - Normality of sample means
- Slightly robust to skew, not to outliers

- Non-parametric tests
  - Use with large outliers or heavy skew
  - Rank sum tests

t-test (Two-Sample)

- Statistical Hypotheses (Two sided)
  - H_0: $\mu_1 = \mu_2$
  - H_a: $\mu_1 \neq \mu_2$

- Difference between two sample means (is it real?)

t-test (Two-Sample)

- $$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Degrees of freedom
  - Conservative is smaller of n_1-1 or n_2-1
  - Software will make less conservative approx.

ANOVA: ANalysis Of VAriance

- One-way ANOVA tests means of more than two samples
- Allows for one test (Less Type 1 error)
- Assumptions:
  - Independant Observations
  - Normality (Watch for outliers)
  - Equal Variances (because we pool sample variances)

ANOVA: ANalysis Of VAriance

- How the test works!
  - Takes <u>between</u> group variation
- Looks at difference in each group mean from overall mean
  - Compares to the <u>within</u> group variation
- F-Statistic compares this ratio
  - 1 mean no extra variance is explained by the groups
  - >1 means the groups are starting to explain more

ANOVA: ANalysis Of VAriance

- Statistical Hypotheses
  - Null H_0: All the population means are equal
  - Alt. H_a: Not all are equal

Multiple Comparisons

- Also called Post-hoc tests
  - *Only do if ANOVA is significant*
- Pair-wise comparisons of group means
  - Have to control for Type I error
- Conservation: Bonferroni Correction
  - Divide α by number of comparisons
- More Power: Tukey‘s Comparisons
  - Use Software for this

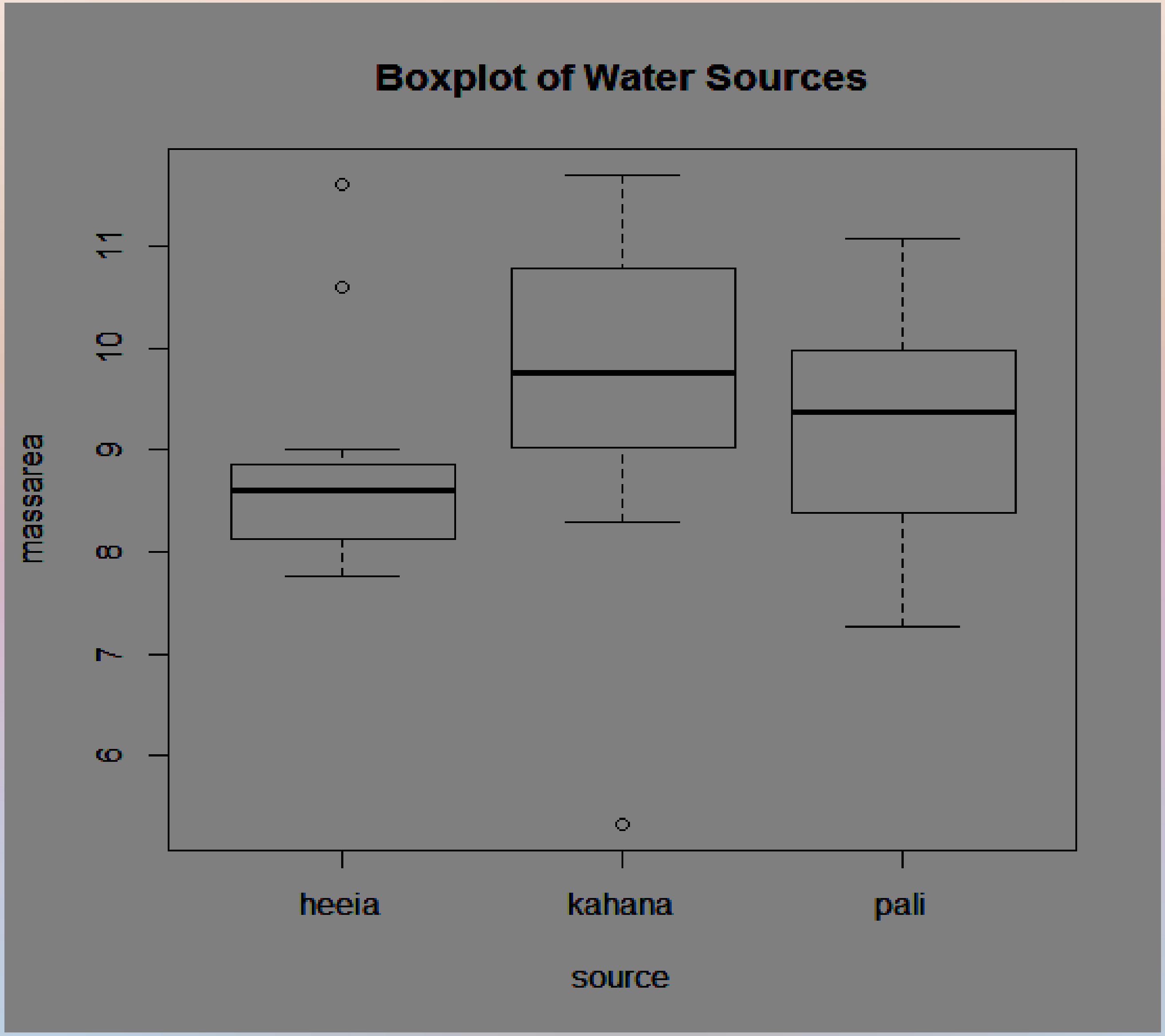
ANOVA

- Terms: source Residuals
- Sum of Squares 2.99195 64.52893
- Deg. of Freedom 2 38

- Residual standard error: 1.303123

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
source	2	2.99	1.496	0.881	0.423
Residuals	38	64.53	1.698		

ANOVA



Boxplot of Water Sources
massarea
6
7
8
9
10
11
heeia
kahana
pali
source

Simple Linear Regression

- Measures change over space and/or time
- Can also be used to measure correlations