Statistical Modeling and Analysis of Scientific Inquiry: The Basics of Hypothesis Testing

So, What is Statistics?

- Theory and techniques for learning from data
  - How to collect
  - How to analyze
  - How to interpret
  - How to present
- Theory and techniques for decision making in the presence of uncertainty

What is Science?

- A process of discovery aimed at understanding the natural world and its inhabitants
  - Observations-Synthesis-Theory-Prediction
  - Repeat as necessary
- Sometimes we call theories “models”
- Models can never truly be validated
  - They can be supported with corroborating evidence
  - They can be demonstrated false

Statistics and Science Go Together

- Statistics provides a toolset that is very useful to science
  - Designing experiments efficiently and effectively
  - Collecting and exploring data
  - Making inferences to investigate synthesizing theories
- Science motivates statistical development
  - Scientific needs drive construction of new statistical tools

Some “Applied Theories”

- We’ve created a new headache medicine: is it better than aspirin?
- Contaminants like heavy metals in soil and water affect plant and animal health
- Yeast respond to cold shock by adjusting gene expression
- Marijuana use alleviates pain in glaucoma patients

Decision making with uncertainty?

Statistics, Uncertainty, Repeatability

- Why do we repeat measurements?
  - Compare alternative treatments, settings, environments
- But why do we repeat even for the same settings?
  - Uncertainty in the measurement
  - Variability in the “experimental unit” (object being studied with measurements

Statistical Modeling In 3 Basic Steps

- Data comprises quantitative measurements of individuals
- Individuals are representative sample from a population
- Population is modeled by a probability density function representing the likelihood of measurement values

Models and Measurements

- The **POPULATION**:
  - A theoretical construct of all possible experimental units that could be studied
  - How we describe or quantify the population is the MODEL
  - PARAMETERS are numbers (*often not known*)that we use to quantify and characterize the **POPULATION**
- The **SAMPLE**:
  - A collection of actual experimental units that are to be studied
  - STATISTICS are numbers that are computed from measurements of the **SAMPLE**

More On Models and Measurements

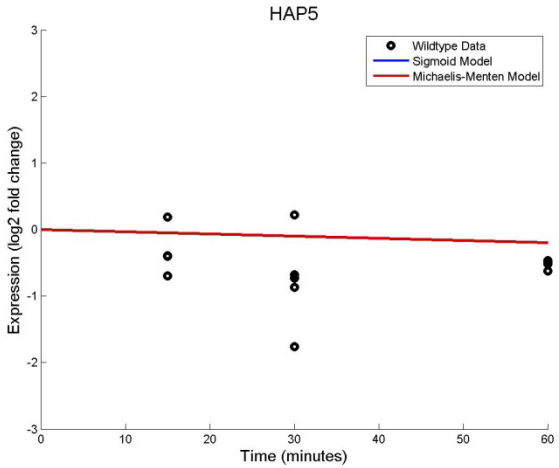
- It is often the case that the POPULATION can not be investigated in its entirety
- SAMPLEs are chosen to be “representative”
  - That is, the variability and behaviors of the POPULATION are captured by the SAMPLE
- Random Sampling is choosing a SAMPLE from the POPULATION so that any POPULATION member is equally likely to be chosen

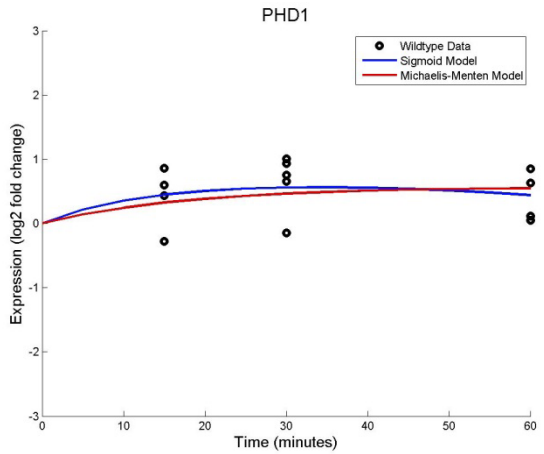
How Does Prior Work Fit?

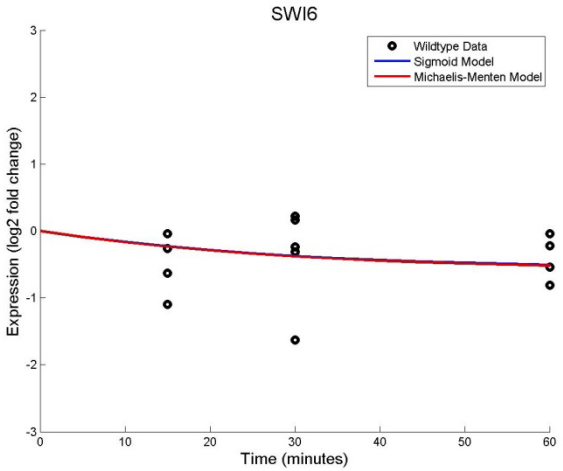
- **Data = Model + Noise**
- Model is a “mean” trend – this is what we’ve considered up to now
- Noise is variability that is (as of now) unaccounted for
- Probability and statistics are fields of inquiry that focus on variability

Anticipation: Gene Expression Models and Data

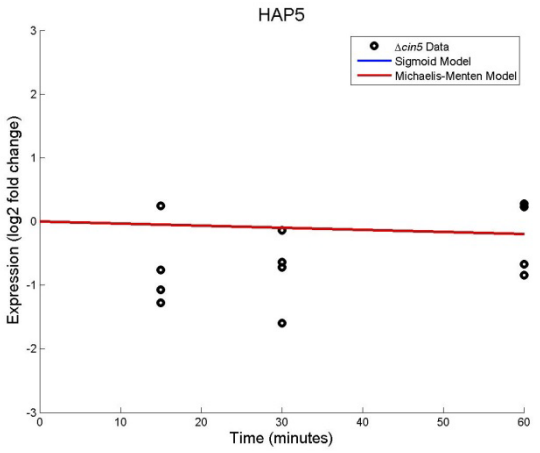
Simulations with wildtype data

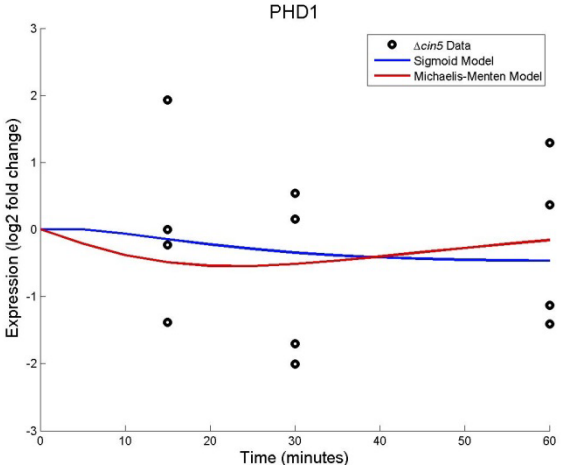


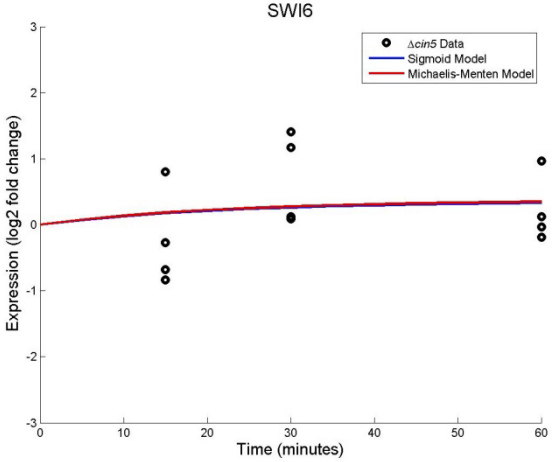




Simulations with Δ*cin5* data







Basic Data Analysis Tools

- Data: $x_1, x_2, \cdots, x_n$
- Mean and median: what's the middle
  - Sample mean, $\bar{x}$,is the average
  - Median is the middle data point (of the sorted list)
- Standard deviation, IQR, median absolute deviation: how much variability

$$s = \sqrt{\frac{1}{n-1}\sum \left(x_i - \bar{x}\right)^2}$$

$$IQR = Q_3 - Q_1$$

$$MAD = median(\left|x - M\right|)$$

- Histograms and box plots: what does the distribution look like?

Our First Statistics

- Given a sample of (scale data) numerical values $x_1, x_2, x_3, \cdots, x_n$
- We compute the sample mean

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + x_3 + \cdots + x_n\right) = \frac{1}{n}\sum x$$

- This number is the center of mass of the data
- Represents “the middle”

Our First Statistics

- Given a sample of (scale data) numerical values $x_1, x_2, x_3, \cdots, x_n$
- We compute the sample standard deviation

$$\bar{x} = \sqrt{\frac{1}{n-1}\left(\left(x_1 - \bar{x}\right)^2 + \left(x_2 - \bar{x}\right)^2 + \cdots + \left(x_n - \bar{x}\right)^2\right)}$$

$$= \sqrt{\frac{1}{n-1}\sum\left(x - \bar{x}\right)^2}$$

- Measures the spread/variability/uncertainty
- Average distance of data points to mean

Order Statistics

- Given a sample $x_1, x_2, x_3, \cdots, x_n$
- We sort them:

$$x^{(1)} \leq x^{(2)} \leq x^{(3)} \leq \cdots \leq x^{(n)}$$

- Median: $m = x^{(n/2)}$
- First and Third Quartiles

$$Q_1 = x^{(n/4)}, Q_3 = x^{(3n/4)}$$

- IQR: $Q_3 - Q_1$

Percentiles

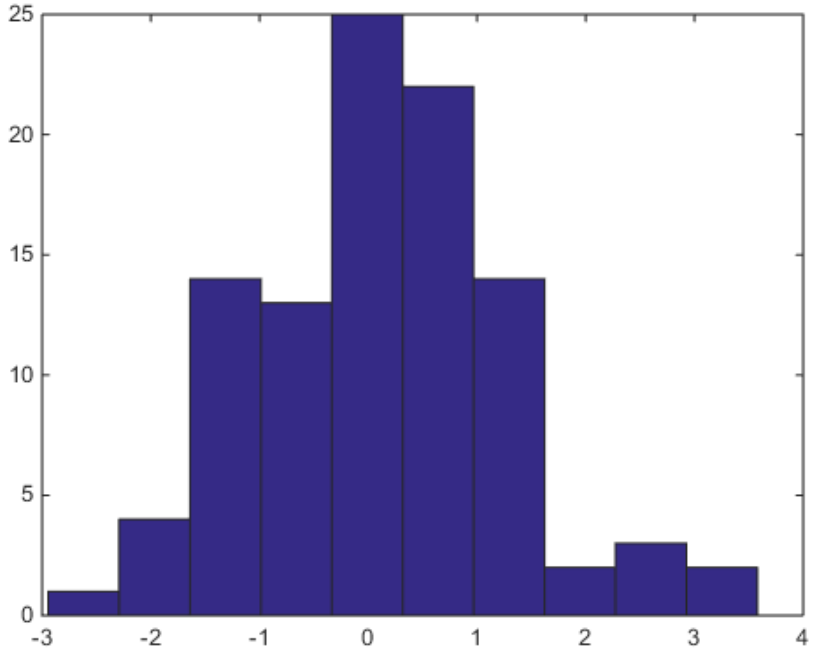
- Given a sample $x_1, x_2, x_3, \cdots, x_n$
- We sort them:

$$x^{(1)} \leq x^{(2)} \leq x^{(3)} \leq \cdots \leq x^{(n)}$$

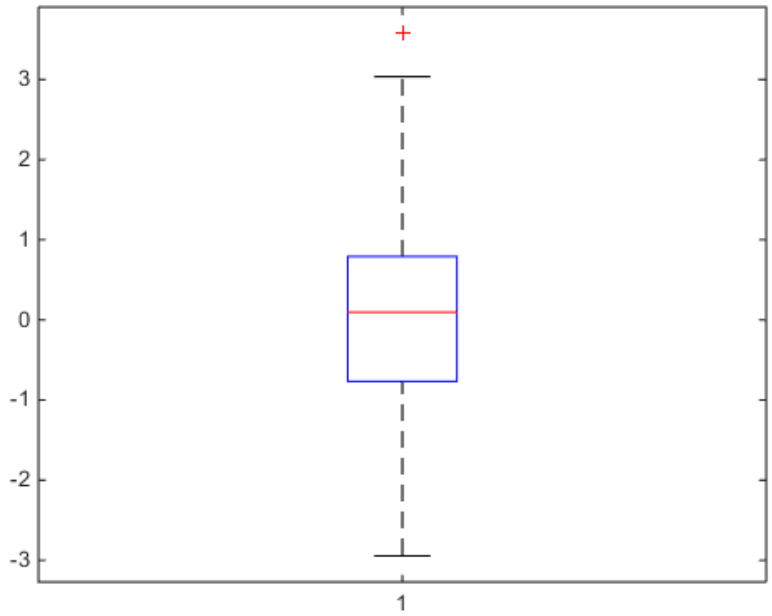
- Percentiles

$$\pi_p = x^{(n*p/100)}$$

Histograms and Box Plots



Each bar is the number of data points between the ordinate values of the bar
Should look like a piecewise constant approximation (like Riemann sums in calc)



The box is bounded by the first and third quartiles, with the mid line being the median.
The whiskers go out to q1-1.5*IQR and q3+1.5*IQR
Outliers are plotted beyond the whiskers

Science and Statistics: An Abstract View

- Theory: we have a population of individuals or "experimental units" (EUs)
  - In bio applications, these are typically organisms
  - In medical applications, these are typically patients
- Inquiry: we propose hypotheses about the properties of these EUs.
  - How an organism respond to stress
  - How a patient responds to treatment
  - Does one treatment work better than another

Principles of Statistical Modeling

- **Modeling Concept 1**: We can characterize the EUs with a vector of attributes that can be observed
- **Modeling Concept 2**: EUs selected randomly from the population produce attributes according to a probability distribution
- **Modeling Concept 3**: The population's probability distribution is known except for a parameter vector that must be estimated from observations
- **Modeling Concept 4**: "Truth" is defined by this unknown parameter vector.

Elements of a Hypothesis Test

- Sample of data
- Two competing hypotheses: the null and its alternative
- A statistic, which is a function of the data with a known sampling distribution
- A rejection criterion against which we assess the statistic's value to decide whether or not we can reject the null.

The Math of Statistics, 1

- The parametrically modeled probability distribution

$$f(x;\theta) = \text{probability density function}$$

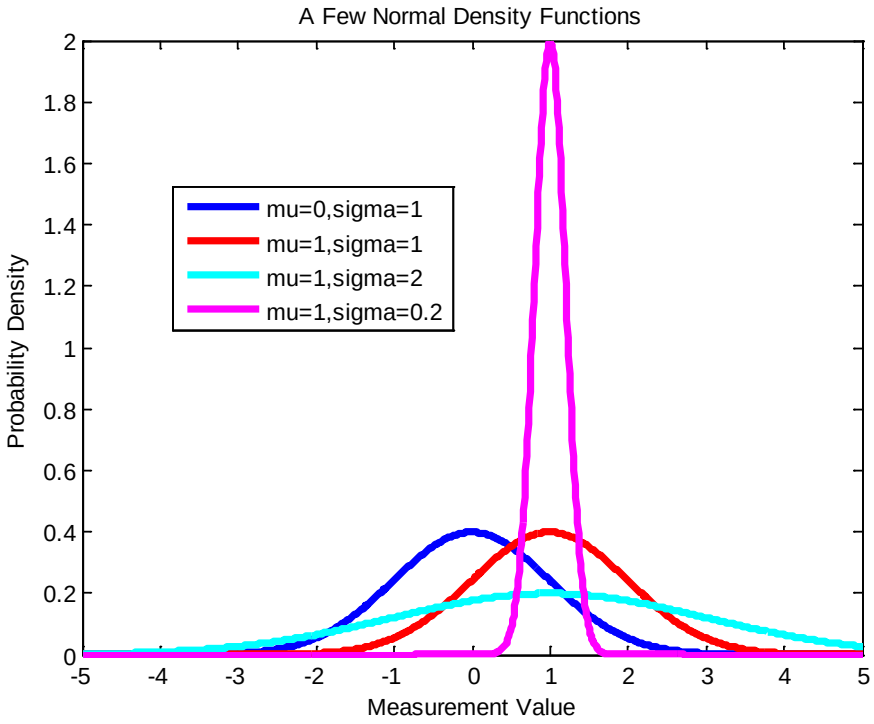
$$x = \text{possible value of the EU observable}$$

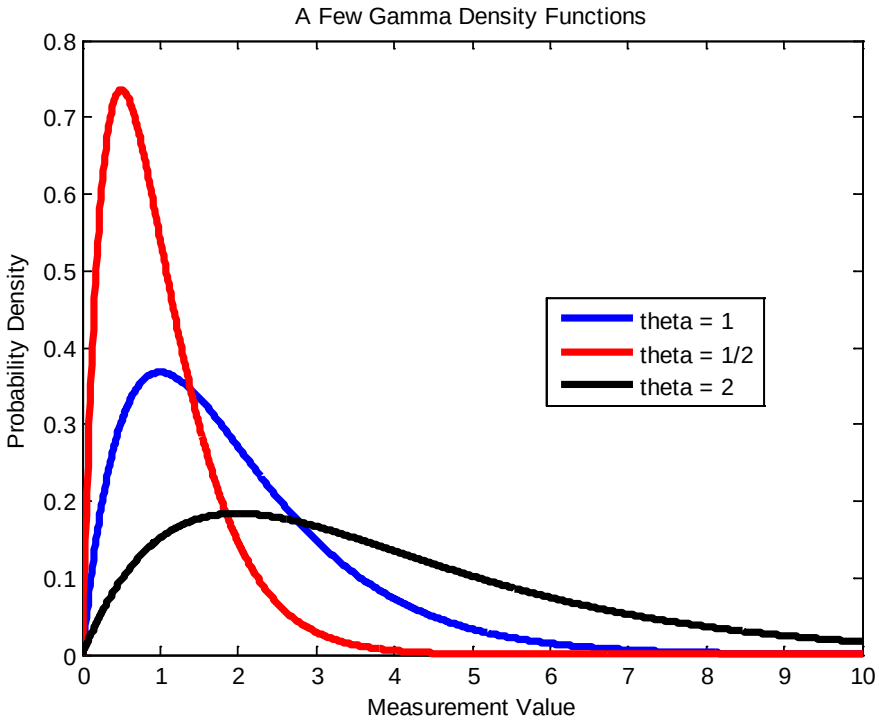
$$\theta = \text{(unknown) parameter characterizing the population}$$

- The parameter θ represents truth about the population
- Question: what can we say about θ after we've seen some x's?

The Math of Statistics, 2

- The probability density models EUs by weighting the possible measurement values





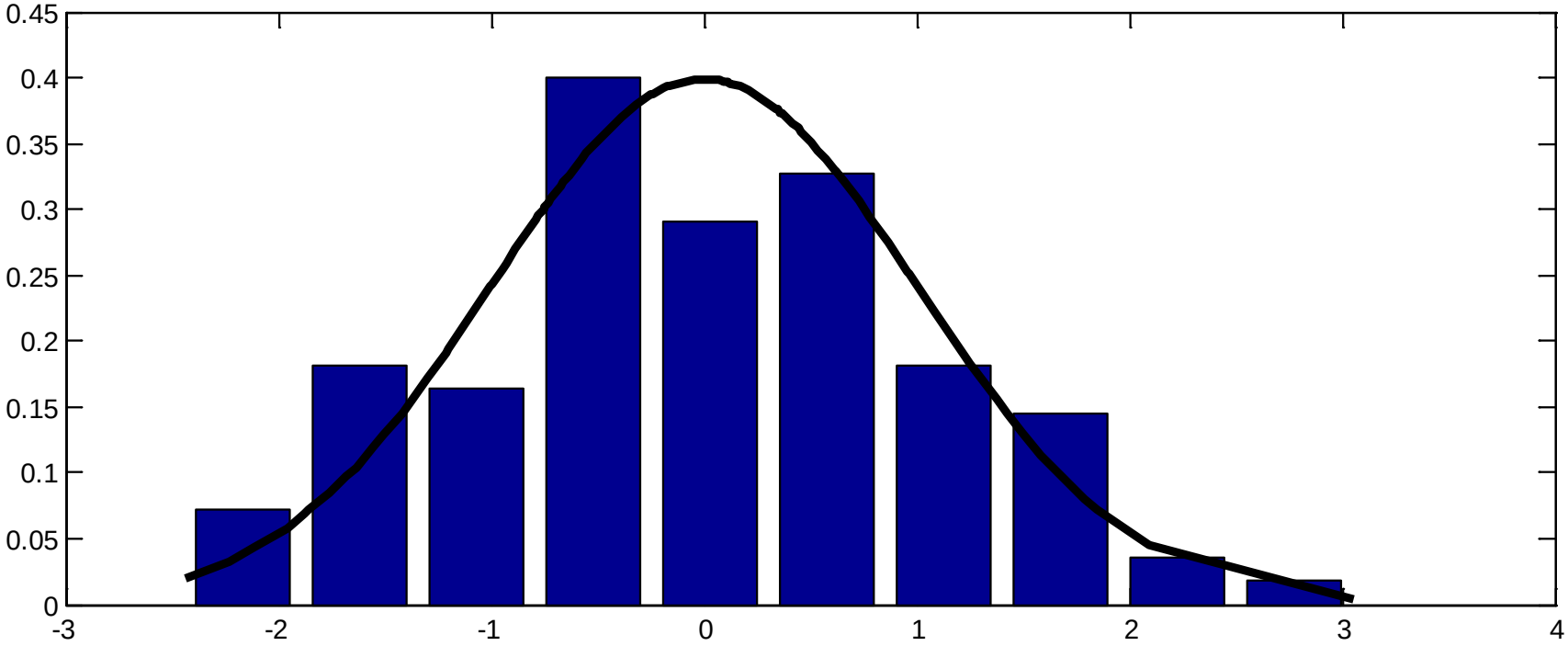
- Area under curve tells us probabilities

The Math of Statistics, 3

- The sample is a collection

$$x_1, x_2, x_3, \cdots, x_n$$

- Ideally the histogram of these would look like the probability density (if we knew θ)



Population vs. Sample

- Population is fixed
  - Very large
  - Impractical to investigate all members
- Population has one distribution
- Population has parameters
  - Fixed, but usually not known

- Samples are random
  - Large enough to be representative
  - Small enough to be studied
- Each sample has a histogram
- Sample has statistics
  - Known, but repeated samples will have different values

Meta: we can think of a population of possible statistic values!!!!!

The biggest idea in statistics

- *In most circumstances*, a larger sample produces an average that more accurately represents a population mean.
- If $x_1, x_2, x_3, \cdots, x_n$ has average $\overline{x}_n$
- If the population has mean μ and std dev σ
- ***Then the population of averages has* mean μ *and* std dev $\sigma/\sqrt{n}$**
- ***And the sample average tends to be normally distributed as n grows***

Hypothesis Testing For the Mean

- Population is characterized by a central value μ and a spread σ of values around that.
  - Should be symmetric
  - Tails should taper relatively quickly
  - The actual values of μ and σ are not known
- The question is the following: Is the unknown μ equal to a specified value μ_0?
  - H_0: $\mu = \mu_0$
  - H_A: $\mu \neq \mu_0$

Mistakes That Can Be Made, 1

- Choosing H_A when H_0 is true
  - Type I error
  - The greek letter α is used denote the likelihood
  - In applications, this is usually **a false positive** or **false detection**.
  - Common approach is to select a value of α we're willing to tolerate
    - α =0.05 is the most common choice
  - Concept: Over many many repetitions when H_0 is true, α percent of the time, we'd declare H_0 to be false

Mistakes That Can Be Made, 2

- Choosing H_0 when H_0 is false
  - Type II error
  - The greek letter β is used denote the likelihood
  - In applications, this is usually **a false negative** or **missed detection**.
  - Common approach is to hope β is small
  - 1 - β is called the **power** of the test
    - Represents the likelihood of detecting a real effect!!!
    - This is the probability of selecting H_A when H_A is true
  - Note that H_A being true is complicated: as long as $\mu \neq \mu_0$ the alternative H_A is true! Even if by 10^{-15} !!!!

Some Concepts and Lingo

- Generally H_0 is something you expect ***not to be true***.
  - For example, you expect a non-zero mean
- In science, models can only be demonstrated to be false.
- We reject an actually true H_0 fairly infrequently (depends on the α we choose)
- When H_0 is not rejected by the test, we say that we "fail to reject H_0," not that we accept H_0.
  - The Type II error probability is difficult to assess

How to Test

- Collect a sample

$$x_1, x_2, x_3, \cdots, x_n$$

- Form the t-statistic

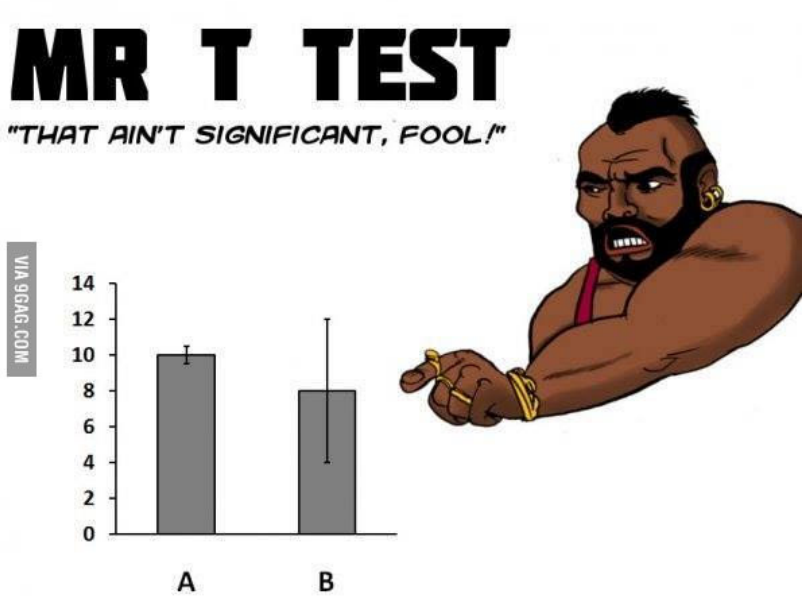
$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \sqrt{n}\,\frac{\bar{x} - \mu_0}{s}$$

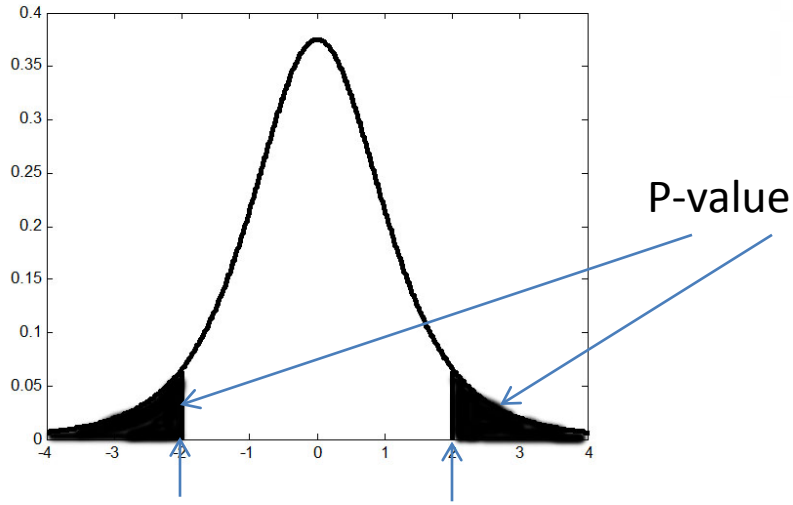
- If H_0 is true, T has a known probability density
  - Student's T distribution with n-1 degrees of freedom
- Choose critical value, t_α, of T distribution
  - Such that $|T| > t_\alpha$ would occur with probability α.

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

The P Value

- Instead of the critical value and the T statistic, we often use α directly with the *p* value statistic
  - Plug the T statistic into its (null) distribution and find the associated probability.

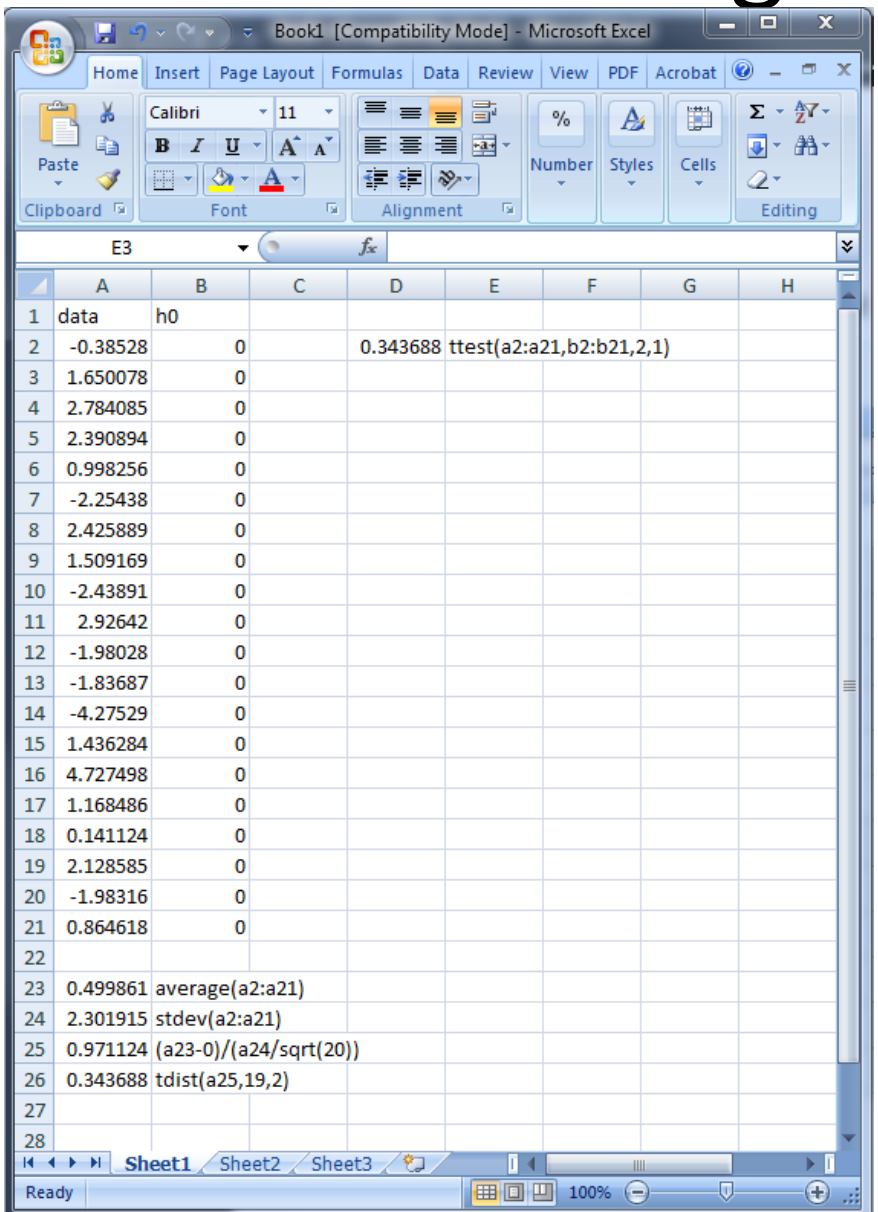




P-value is the shaded area added together

T value and its minus

Doing this in Excel



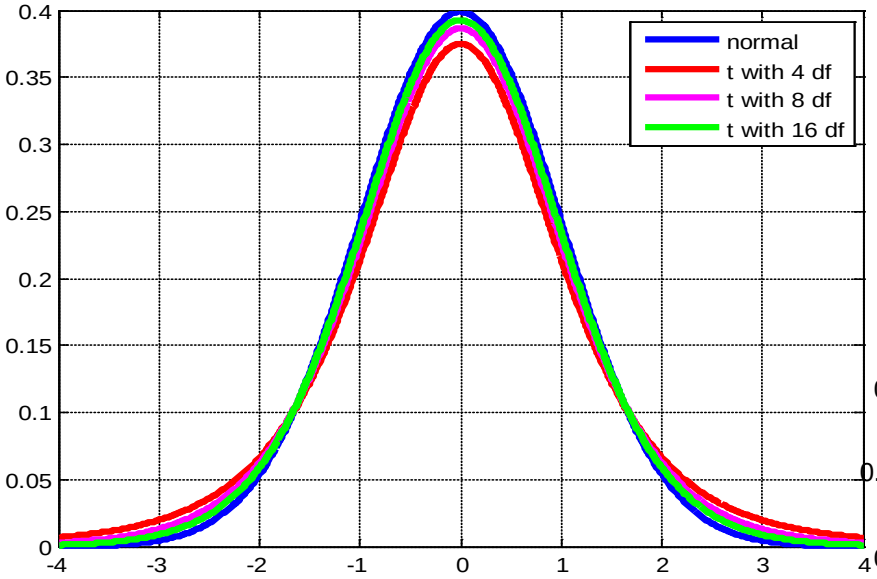
	A	B	C	D	E	F	G	H
1	data	h0						
2	-0.38528	0		0.343688	ttest(a2:a21,b2:b21,2,1)			
3	1.650078	0						
4	2.784085	0						
5	2.390894	0						
6	0.998256	0						
7	-2.25438	0						
8	2.425889	0						
9	1.509169	0						
10	-2.43891	0						
11	2.92642	0						
12	-1.98028	0						
13	-1.83687	0						
14	-4.27529	0						
15	1.436284	0						
16	4.727498	0						
17	1.168486	0						
18	0.141124	0						
19	2.128585	0						
20	-1.98316	0						
21	0.864618	0						
22								
23	0.499861	average(a2:a21)						
24	2.301915	stdev(a2:a21)						
25	0.971124	(a23-0)/(a24/sqrt(20))						
26	0.343688	tdist(a25,19,2)						
27								
28								

- Data in a column or row
- Compute the sample mean with the average function
- Compute the sample standard deviation with the stdev function
- Compute the t statistic

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \sqrt{n}\frac{\bar{x} - \mu_0}{s}$$

- Compute the p-value by plugging the t statistic into the integral with tdist(T,n-1,2)
  - That last 2 is for two-tailed integral
- Alternatively, use ttest to compute.
  - Ttest is designed for two-sample comparison, so you have to trick it by creating a sample with all μ_0's

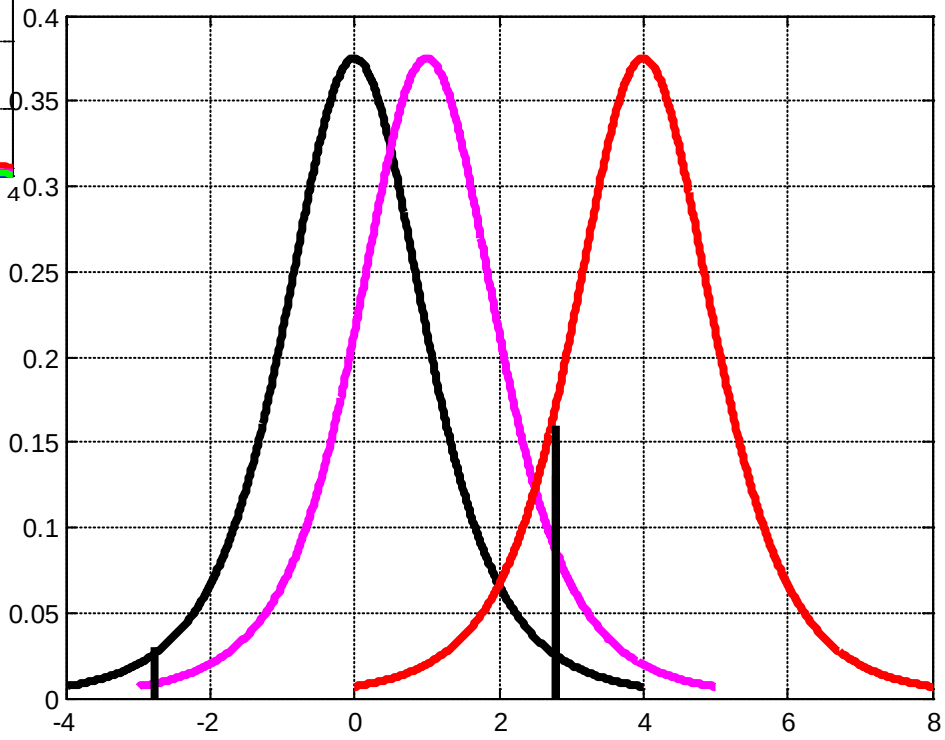
More On Student's T



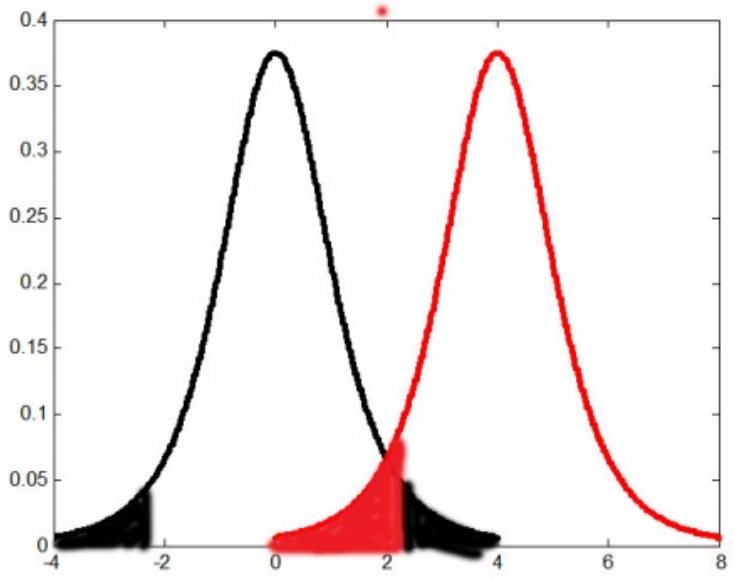
Slightly false null:
Centered near 0

Null true:
Centered at 0

Extremely false null:
Centered far from 0



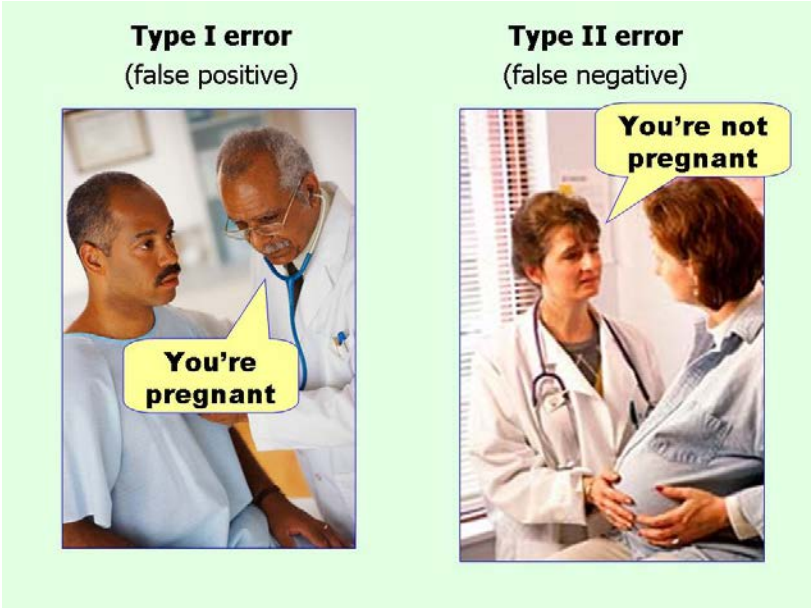
Type I and Type II



α is the black shaded:
Depends Only on Null

β is the red shaded:
Depends on how far the red curve is shifted

Some alternatives are easier to detect



The Alternative Hypothesis

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \sqrt{n}\,\frac{\bar{x} - \mu_0}{s}$$

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

- If H_0 is true, T has Student's T distribution with n-1 degrees of freedom
- If H_A is true, then

$$T' = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \sqrt{n}\,\frac{\bar{x} - \mu}{s}$$

has the T distribution!

The Alternative Hypothesis

$$T' = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\bar{x} - \mu_0 + \mu_0 - \mu}{s/\sqrt{n}}$$

$$= T + \frac{\mu_0 - \mu}{s/\sqrt{n}}$$

$$= T + \frac{d}{s/\sqrt{n}}$$

The Alternative Hypothesis

- We fail to reject the null when

$$T < t_\alpha \Leftrightarrow T' < t_\alpha - \frac{d}{s/\sqrt{n}} = t_\beta^{LT}$$

- What this tell us:
  - If we have s and n fixed, an effect of size *d* leads to a power of 1 – β.
  - If we have s and n fixed, a power of 1 – β requires an effect size no smaller than *d*.
  - If we want a power of 1 – β and an effect size of d, then we need n samples to achieve our goal.

Effect Size, Sample Size, and Power

- To detect an alternative of $d = |\mu - \mu_0|$ with power 1-β, we need

$$n = \frac{s^2}{d^2}\left(t_{\alpha,n-1} + t^{(1)}_{\beta,n-1}\right)$$

- With n samples, an effect size of d can be detected with power from

$$t^{(1)}_{\beta,n-1} = \frac{d}{s/\sqrt{n}} - t_{\alpha,n-1}$$

Multi-Group Similarity Testing

- Population comprises a fixed set of groups: 1,2, …, p
  - Usually thought of as "statistically identical" individuals within the groups
  - Each group receives a different "treatment"
  - Process leads to groups that may have different means $\mu_1,\ldots\ \mu_{p,}$
  - Groups have the same variance σ^2
  - We sample from each group, size $n_1,\ldots n_p$
- The question is the following: Is at least one treatment different?
  - H_0: $\mu_1 = \mu_2 = \ldots\ \mu_p$
  - H_A: At least one of the μ_i's is different

A Digression

- Given two numbers, how do we compare them?
  - Subtract to compute the difference
  - Divide to compute the ratio
- Statistical use of subtraction relies on T-statistics
  - Two numbers are equal if difference is 0
- Statistical use of division relies on F-statistics
  - Two numbers are equal if ratio is 1

Probability Density Functions

- The normal distribution(mu,sigma): bell shaped, with
  - mu+/- sigma containing 68%
  - mu+/- 2sigma containing 95.4%
  - mu+/- 3sigma containing 99.7%
- Chi squared (m)
  - This distribution is what you get when you square m normal(0,1)'s and add them up
  - $\chi = Z_1^2 + Z_2^2 + Z_3^2 + \cdots + Z_m^2$
  - The quantity below is chi squared (n-1)

$$(n-1)\frac{S^2}{\sigma^2} = \left(\frac{x_1 - \bar{x}}{\sigma}\right)^2 + \cdots \left(\frac{x_n - \bar{x}}{\sigma}\right)^2$$

Probability Density Functions

- The T-distribution comes from dividing a normal(0,1) by the square root of a chi-squared

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \sqrt{n}\,\frac{\bar{x} - \mu_0}{s}$$

- The F-distribution comes from a ratio of chi-squareds

</digression>: ANOVA

- Collect a sample

$$x_{11}, x_{12}, x_{13}, \cdots, x_{1n_1} : \text{treatment } 1$$

$$x_{21}, x_{22}, x_{23}, \cdots, x_{2n_2} \text{ treatment } 2$$

$$\vdots$$

$$x_{p1}, x_{p2}, x_{p3}, \cdots, x_{pn_p} \text{ treatment } p$$

- Test the hypothesis:

> H_0: $\mu_1 = \mu_2 = \ldots \mu_p$
> H_A: At least one of the μ_i's is different

- Assumption: common variance σ^2

How To Test

- All treatments have the same mean under H0

$$\bar{x} = \frac{1}{n}\sum_j \sum_i x_{ij} : \text{Grand mean}$$

$$\bar{x}_j = \frac{1}{n_j}\sum_i x_{ij} : \text{Group means}$$

$$S^2_{Full} = \frac{1}{n}\sum_j \sum_i \left(x_{ij} - \bar{x}_j\right)^2$$

$$S^2_{H0} = \frac{1}{n}\sum_j \sum_i \left(x_{ij} - \bar{x}\right)^2$$

$$F = \frac{n-p}{p-1}\frac{S^2_{H0} - S^2_{Full}}{S^2_{Full}} \text{ is } F(p\text{-}1, n\text{-}p)$$

Pseudo-ANOVA

- Collect a sample

$$x_{11}, x_{12}, x_{13}, \cdots, x_{1n_1} : \text{treatment } 1$$

$$x_{21}, x_{22}, x_{23}, \cdots, x_{2n_2} \text{ treatment } 2$$

$$\vdots$$

$$x_{p1}, x_{p2}, x_{p3}, \cdots, x_{pn_p} \text{ treatment } p$$

- Test the hypothesis:

> H_0: $\mu_1 = \mu_2 = \ldots \mu_p = 0$
> H_A: At least one of the μ_i's is different from 0

- Assumption: common variance σ^2

How To Test

- All treatments have the same mean under H0

$$0 : \text{Hypothesized mean}$$

$$\bar{x}_j = \frac{1}{n_j}\sum_i x_{ij} : \text{Group means}$$

$$S^2_{Full} = \frac{1}{n}\sum_j \sum_i \left(x_{ij} - \bar{x}_j\right)^2$$

$$S^2_{H0} = \frac{1}{n}\sum_j \sum_i \left(x_{ij} - 0\right)^2$$

$$F = \frac{n-p}{p}\frac{S^2_{H0} - S^2_{Full}}{S^2_{Full}} \text{ is } F(p, n\text{-}p)$$