McMaster University | Integrated Biomedical Engineering & Health Sciences Program

IBEHS - 4QZ3
Modelling of Biological Systems

Lecture 2

TAYLOR DEVET MASC.
PHD. CANDIDATE BIOLOGICAL AND BIOMEDICAL ENGINEERING
MCGILL UNIVERSITY
SHRINERS HOSPITAL FOR CHILDREN

Todays Aims…

Statistics Review

Linear Regression

Experimental Design

Reminders

Quizzes

- Friday – Sunday

Group Project

- Let Noor and Andrew know your groups so we can sort remaining people into groups

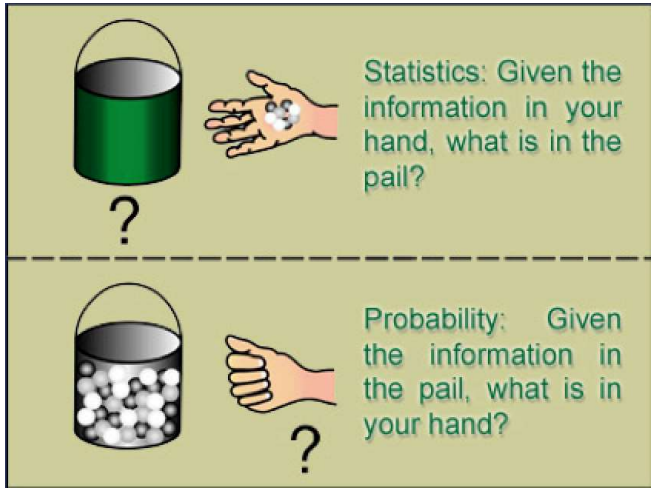
Assignment 1 due October 3rd

Statistics Review

Math
- study of space, change, structure, quantity
- Science or order, structure and relationships
- Given data, make model

Statistics
- Collection, analysis, explanation, interpretation of data
- Type of mathematical science
- Given data, predict model

Generally
- Determine population to look at OR model to study
- Collect data using survey or experiment
- Analyze data to look for significance



https://mathprojects.com/tag/statistics/

Math vs. Statistics

Math	Statistics
Calculations and results are **independent of context**	**Context is paramount** from input of data to the final result.
Measurements are taken of **specific attributes** (like length or weight)	Measurements are usually **challenging to measure**, e.g. aptitude, pace of life, disease progressions.
Rigorous **deductive reasoning** usually leads to **a "correct" answer**.	**Inductive reasoning** leads to **uncertain conclusions.**
A **few words** in everyday use have technical meanings. For example "factor" and "multiple".	A **plethora** of everyday words have very specific meanings, i.e.: bias, accuracy, normal, confidence and significance.

https://storage.ning.com/topology/rest/1.0/file/get/6446165888?profile=original

Branches of Statistics

Descriptive Statistics

- Summarize data using standard deviation, mean, median, mode etc
- Look at shape of data, skewness, kurtosis etc
- Focuses on learning about a sample rather than population
- Generally used for non parametric data
- Take large amount of data and describe it using parameters

Inferential Statistics

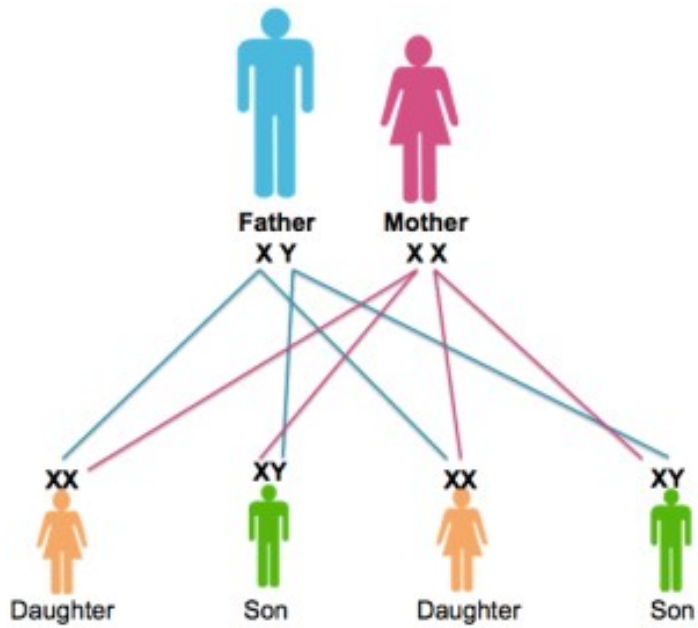
- Make conclusions about data that have random variance
- Uses data to make predictions
- I.e. use sample mean to make inferences about population mean
- Hypothesis testing to answer research question

Probability

Branch of mathematics

Math to describe how likely something is to happen

How probable it is that a statement is true



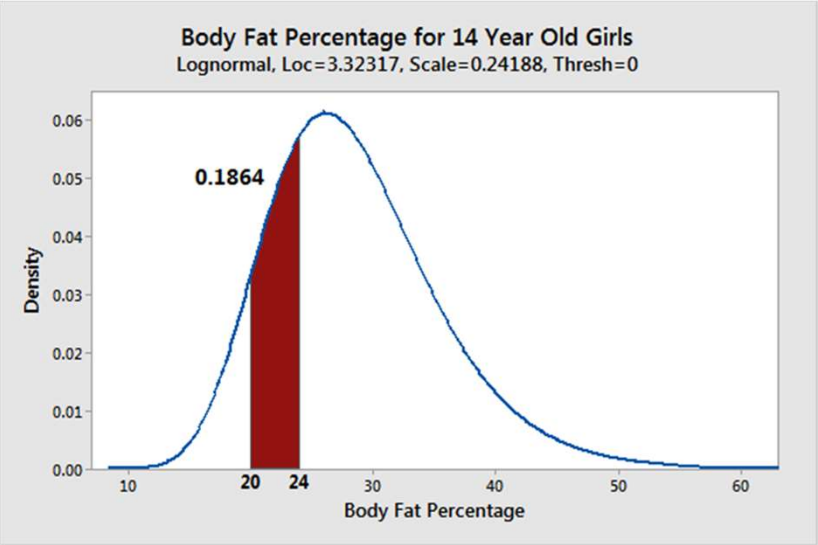
https://allinonehighschool.com/using-punnett-squares-to-predict-offspring/

Probability Distributions

Display the **likelihood** of obtaining a value given **all the possibilities** of a random variable

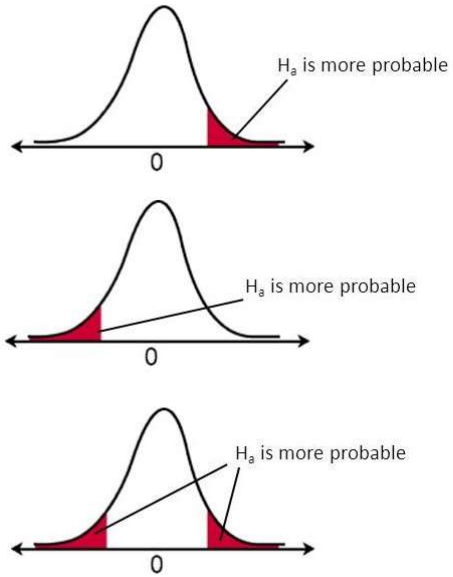
P(x)

- function that shows the likelihood that a random variable will be the specific value of x



Hypothesis Testing

- Hypothesis tests look at 2 mutually exclusive statements regarding the population and determine which is more true
- Null Hypothesis (Ho)
  - No difference or effect
  - Accepting this leads to no change
- Alternate Hypothesis (Ha)
  - There is some difference or effect
  - Accepting this leads to a change
- Critical region
  - Region of value that corresponds to rejection of Ho



Right-tail test

H_a: $\mu > \text{value}$

Left-tail test

H_a: $\mu < \text{value}$

Two-tail test

H_a: $\mu \neq \text{value}$

https://towardsdatascience.com/everything-you-need-to-know-about-hypothesis-testing-part-i-4de9abebbc8a

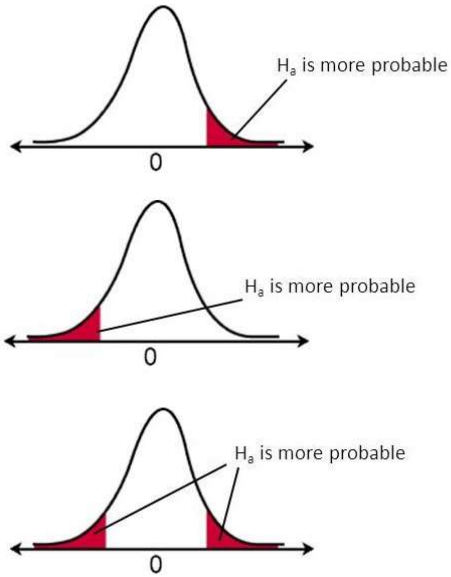
Hypothesis Testing

One tailed Test

- Critical area is one sided
- Critical area is either greater or less than critical value, not both
- If sample falls into area, Ha is accepted

Two Tailed Test

- Critical area is two sided
- Critical area is greater than or less than critical values
- If sample falls into either of the critical areas, Ha is accepted



Right-tail test

H_a: $\mu >$ value

Left-tail test

H_a: $\mu <$ value

Two-tail test

H_a: $\mu \neq$ value

https://towardsdatascience.com/everything-you-need-to-know-about-hypothesis-testing-part-i-4de9abebbc8a

Steps in Hypothesis Testing

1. Formulate Hypothesis
2. Select Test type
3. Pick significance level
4. Collect data
5. Determine Critical value of test statistic
6. Determine if Test statistic falls into rejection region
7. Reject or don't reject Ho

Test Statistic

A measure of how close the sample comes to the null hypothesis

Gives information relevant to deciding if Ho should be rejected

Each distribution uses its own test statistic

- Z test – Z Statistic
- T-test – t-statistic
- ANOVA – F Statistic
- Chi-square – Chi-Square statistic

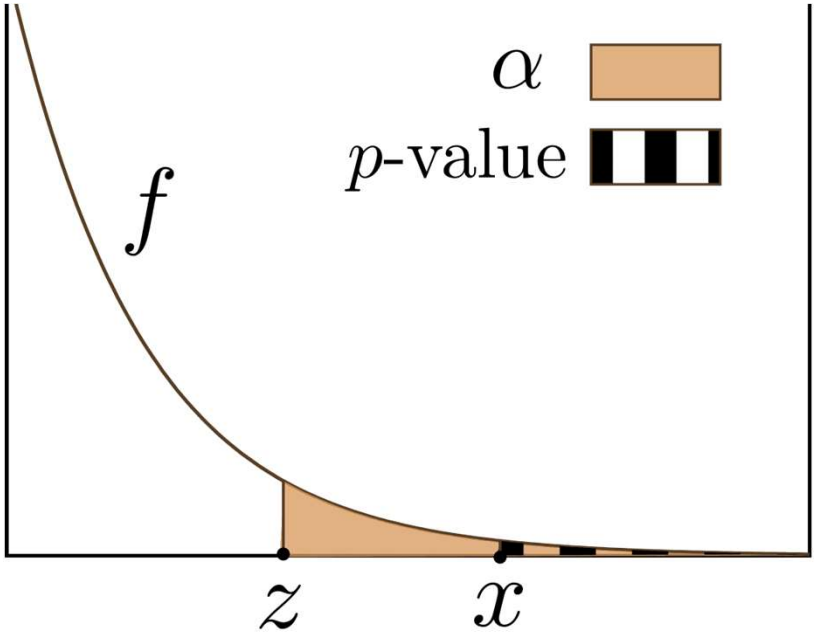
P Value

The probability of getting the result *at least* as extreme as the observed results

Assumes the null hypothesis is correct

Alpha value

- Significance Level
- If p < alpha, accept the null hypothesis



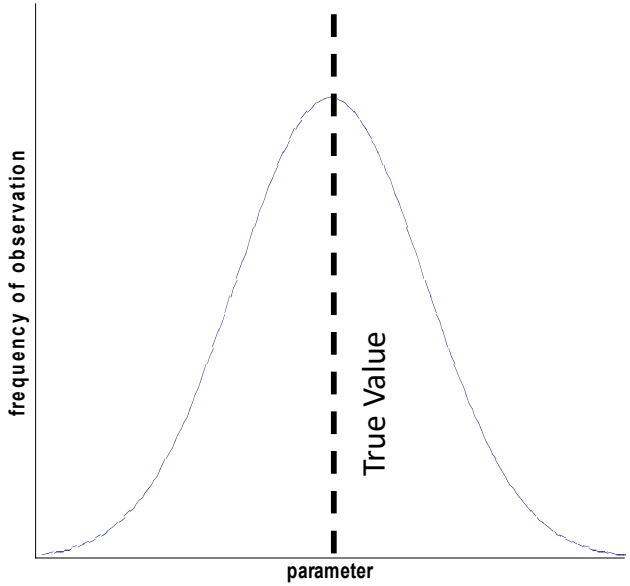
https://www.investopedia.com/terms/p/p-value.asp

Summarizing Data

Need to differentiate between population and sample

Population: infinite data, looks at entire population

Sample data: looks at a sub section of the population



Parameters to describe distribution

Gaussian distribution

- Mean (~true value)
- Variance (~variability)

Non-Gaussian distribution

- Median (~true value)
- at least 2 Percentiles (~variability)

Apr. 30, 1777 in Brunswick, Died: Feb 23, 1855 in Gottingen

Interpretation

In a balanced (gaussian) distribution

- Mean == true value
- Variance or standard deviation characterize uncertainty in the individual measurement
- ~68% of measurements are within 1σ
- ~95% of measurements are within 2σ

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right)$$

Likelihood/frequency of getting a value for a measurement

Mean

Population Mean:

$$\lim_{N\to\infty}\sum_{i=1}^{N}\frac{x_i}{N}=\mu$$

Sample Mean:

$$\bar{x}=(x_1+x_2+x_3+\ldots+x_n)/n$$

Arithmetic Mean:

$$\bar{x}=\frac{1}{n}\cdot\sum_{1}^{n}x_n$$

AKA: average, centroid, <x>, etc.

Measurement of Variability

- variance and standard deviation

Population variance:

$$\sigma^2 = \lim_{n\to\infty}\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{N}$$

$$= \lim_{n\to\infty}\sum_{i=1}^{N}\frac{x_i^2}{N} - \left[\lim_{n\to\infty}\sum_{i=1}^{N}\frac{x_i}{N}\right]^2$$

Population Standard Deviation:

$$\sigma = \sqrt{\sigma^2}$$

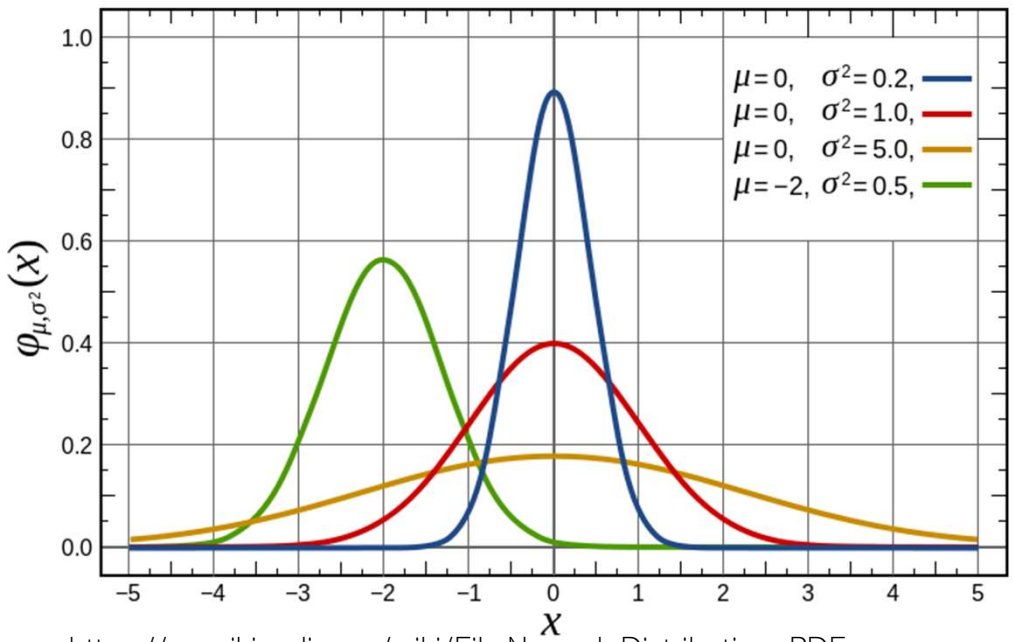
Sample Variance:

$$s^2 = \sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{n-1}$$

Sample Standard deviation

$$s = \sqrt{s^2}$$

Shapes of Normal Curves

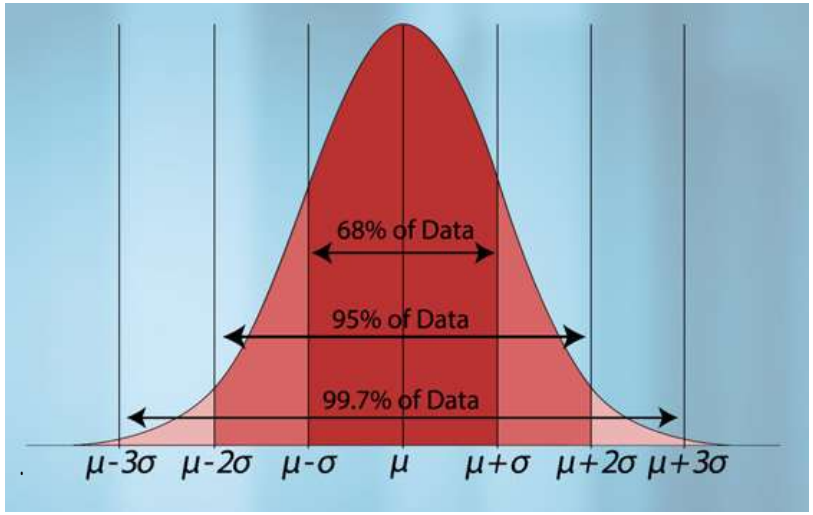


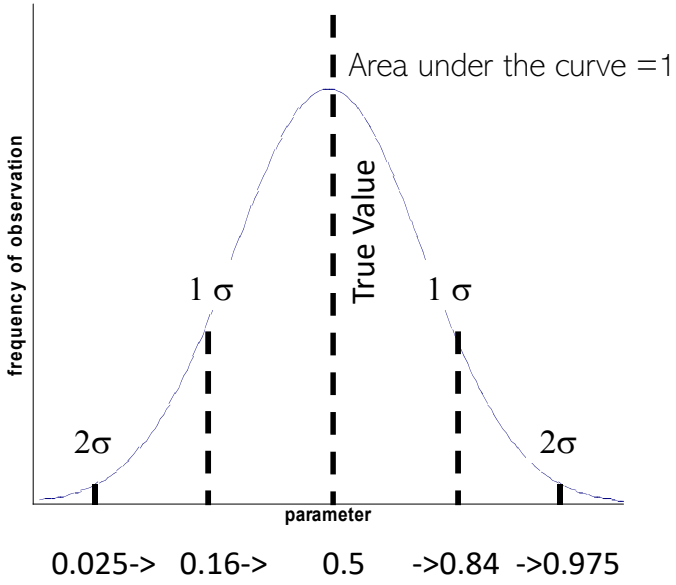
https://en.wikipedia.org/wiki/File:Normal_Distribution_PDF.svg

Z-Distribution vs Normal Distribution

Most test assume/require the data to be normal distributed

Z is a special case with mean=0 and SD = 1





Estimating Population Parameters

e.g if N < infinite

Sampling n measurements

estimating variance s^2 and mean $\bar{x}$ from a limited number of samples

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

<u>*OR*</u>:

$$s^2 = \frac{1}{n-1} \cdot \left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^{n} x_i \right)^2 \right]$$

Central Limit Theorem

If we were able to measure repeated $\bar{x}$ ➔ then we would approach μ

(i.e. mean of all experiments is equal to the population mean)

AND:

Therefore, the larger we can make our sample the closer we approach the population mean

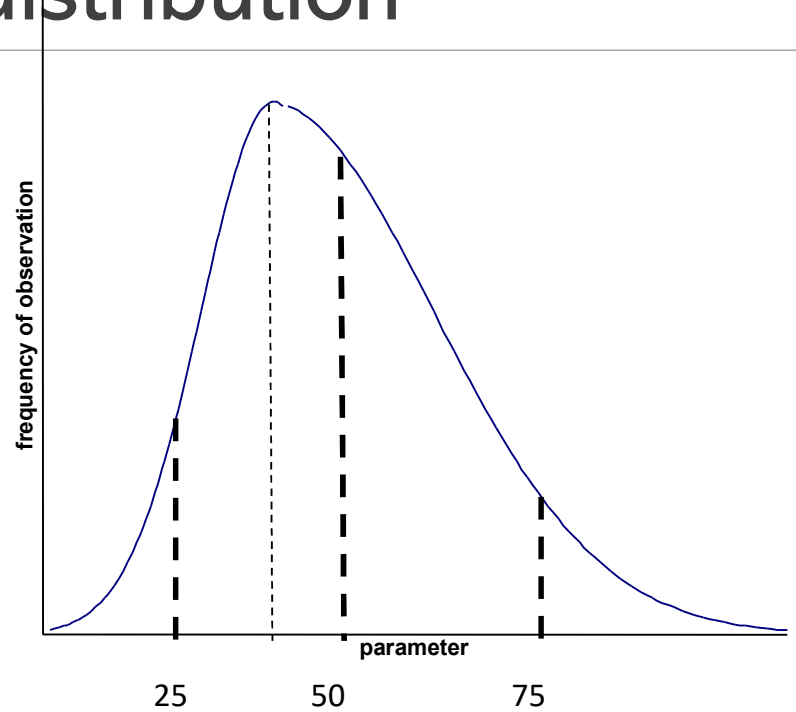
Measure of variability of a sample

Sample variability is described by the standard error or SEM

$$\boxed{SEM = \sigma / \sqrt{n}}$$

- No matter what the initial distribution of x_i are, as n is getting larger the distribution of x_i will approach a normal distribution
- The distribution of $\sum_{i=1}^{n} \overline{x_i}$ will approach μ

Skewed distribution



Skewness

- measure of the asymmetry of data around the sample mean.
- If skewness is –ve, data are spread out more to the left of the mean than to the right.
- If skewness is +ve, data are spread out more to the right.
- The skewness of the normal distribution (or any perfectly symmetric distribution) is zero.
- (AKA 3rd moment about the mean)

$$m_3 = \frac{\sum\left(X-\overline{X}\right)^3}{n} \qquad m_2 = \frac{\sum\left(X-\overline{X}\right)^2}{n}$$

Kurtosis

- a measure of how outlier-prone a distribution is.
- The kurtosis of the normal distribution is 3.
- Distributions that are more outlier-prone than the normal distribution have kurtosis greater than 3
- distributions that are less outlier-prone have kurtosis less than 3.
- (AKA 4th order moment about the mean)

$$m_4 = \frac{\sum\left(X - \overline{X}\right)^4}{n}$$

- sometimes kurtosis-3 is presented so the distribution is around 0.

Example

1.1650	
0.6268	$m_4 = 0.3131$
0.0751	$m_3 = -0.688$
0.3516	$m_2 = 0.3802$
-0.6965	

Tests for Normality

1. Jarque-Bera
   - evaluates the hypothesis that x has a normal distribution with unspecified mean and variance
   - vs the alternative that x does not have a normal distribution.
   - based on sample skewness (s) and kurtosis (k) of n samples of x.
   - tests whether the sample skewness and kurtosis are unusually different than their expected values
   - should not be used with small samples.

$$JB = \frac{n}{6}\left(s^2 + \frac{(k-3)}{4}\right)$$

2. Lilliefors
   - useful for smaller samples
   - similar to the Kolmogorov-Smirnov test
   - Looks at Cumulative Distribution function – the probability that x is $\leq \bar{x}$

Other distributions

Not all data is "Normal" (i.e. Gaussian distribution)

Binominal
- head/tail coin flipping

Lorentzian
- resonance in NMR

Poisson
- radioactive decay

Normality should always be assessed.
- histogram analysis
- Kurtosis - distribution has longer tails than normal
- Skewness - data not distributed evenly about a mean

Binomial Distribution

- Parameters n and p that represent a boolean question
- Probability of a number of "yes" in a row
- Is often drawn with a continuous curve but is discrete

The probability $P_B(x;n,p)$ for observing x of n items to be in the state with probability p is given by the binomial distribution:

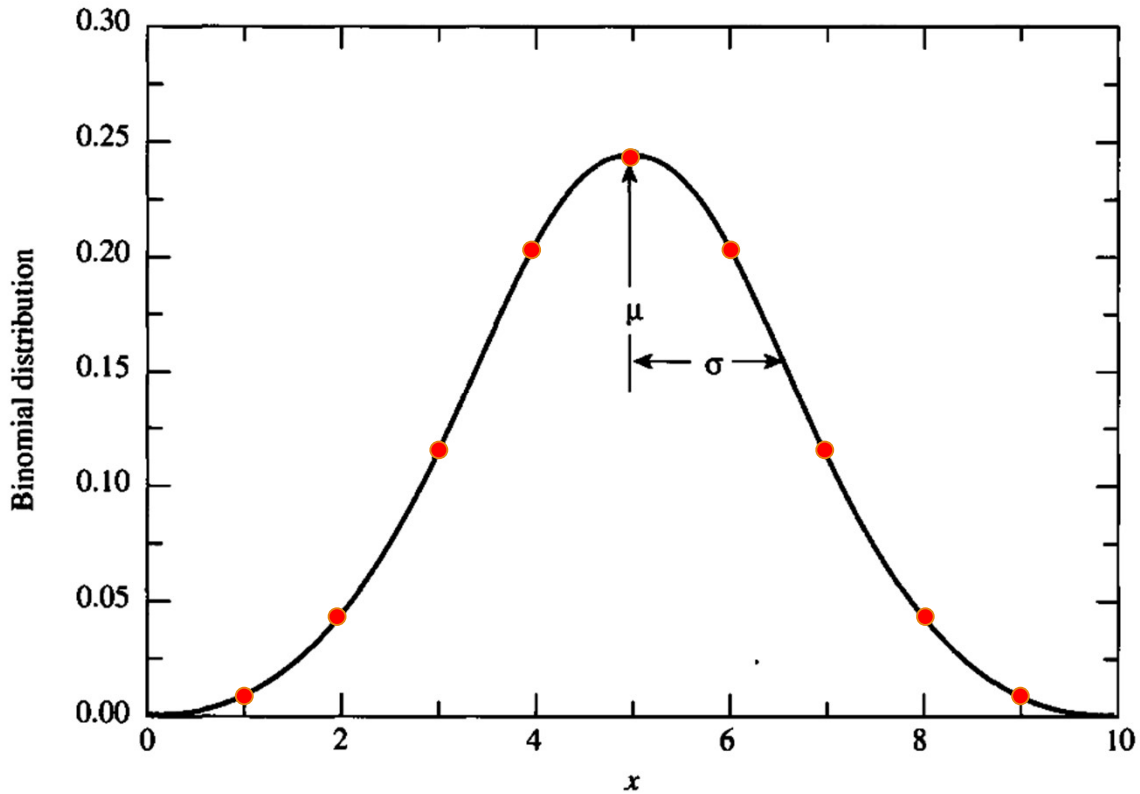
$$Px = \sum_{k=0}^{n} \binom{n}{x} p^x q^{n-x}$$

Binomial Distribution

$$P_B(x; n, p) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

$$\mu = \sum_{x=0}^{n} \left[x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \right] = np$$

$$\sigma^2 = \sum_{x=0}^{n} \left[(x-\mu)^2 \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \right] = np(1-p)$$



Binomial Distribution for μ=5.0 and p=0.5. The curve is shown as continuous but in reality the function is only defined as the discrete points (red dots)

Example

An engineer working at a particle accelerator makes preliminary measurements of the angular distribution of K mesons scattered from a $H_2(l)$ target. They know there should be equal numbers of particles scattered forwards and backwards in the centre-of-mass system of particles. She measures 1000 interactions and finds 472 scatter forwards and 528 backwards. What uncertainty should be quoted?

For uncertainty use Standard Deviation:

Poisson Distribution

Probability of observing x events in a set period of time t if events occur with a known constant mean and are independent of previous events

Also a discrete distribution but is represented by a continuous curve

$$P_P(x; \mu) = \frac{\mu^x}{x!} e^{-\mu}$$

μ = expected value of x, positive real number, mean number of events over time t

x = number of occurrences

Poisson Distribution

$$P_P(x;\mu) = \frac{\mu^x}{x!}e^{-\mu}$$

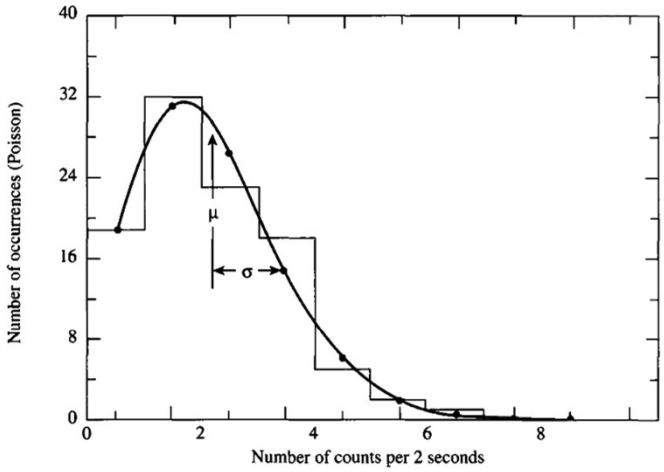
$$\langle x \rangle = \sum_{x=0}^{\infty}\left(x\frac{\mu^x}{x!}e^{-\mu}\right) = \mu e^{-\mu}\sum_{x=1}^{\infty}\frac{\mu^{x-1}}{(x-1)!} = \mu e^{-\mu}\sum_{y=0}^{\infty}\frac{\mu^y}{y!} = \mu$$

$$\sigma^2 = \langle (x-\mu)^2 \rangle = \sum_{x=0}^{\infty}\left[(x-\mu)^2\frac{\mu^x}{x!}e^{-\mu}\right] = \mu$$

Therefore, the standard deviation, σ is equal to the square root of the mean, m and the Poisson distribution has only a single parameter, m.

Example

In an experiment to determine mean life of radioactive isotopes of silver, a grad student detected background counts from cosmic rays. Values were recorded as counts on their detector for a series of 100, 2-second intervals and the mean number of counts was found to be 1.69 per interval. Using the mean they estimated the standard deviation to be:



<u>Notes</u>

1) Poisson is defined as discrete points but here shown as a continuous curve

2) As m increases the symmetry of the Poisson distribution increases until it becomes indistinguishable from a Gaussian

Lorentzian Distribution

Also known as Cauchy distribution

Continuous distribution

Lorentzian Probability Density Function
$P_L(x; \mu, \Gamma)$

- appropriate for data exhibiting resonant behavior

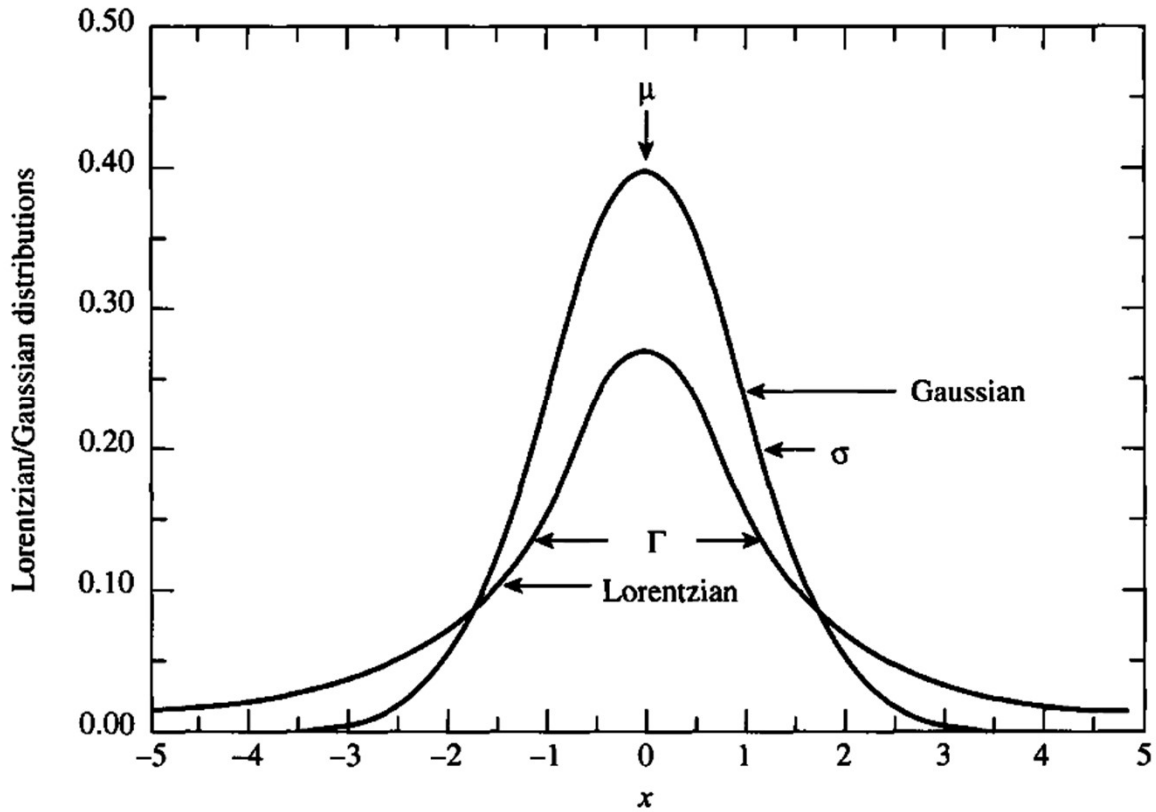
- mean μ, and full width at half maximum (FWHM), Γ

- similar to Gaussian, but doesn't diminish to zero as fast

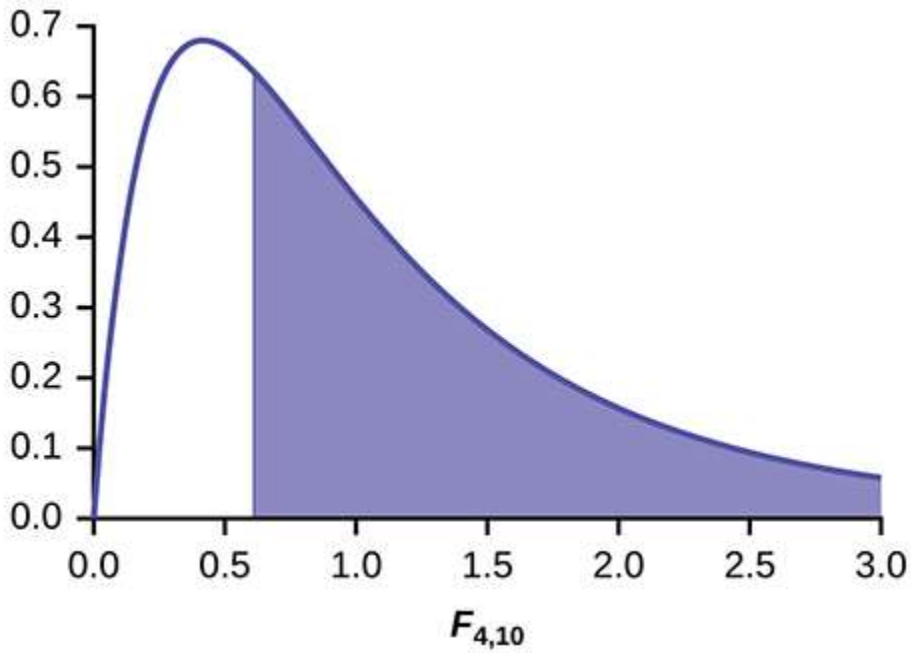
- undefined mean and variance

$$p_L(x; \mu, \Gamma) = \frac{1}{\pi} \frac{\Gamma/2}{(x-\mu)^2 + (\Gamma/2)^2}$$

$$\sigma^2 = \langle (x-\mu)^2 \rangle = \frac{1}{\pi} \frac{\Gamma^2}{4} \int_{-\infty}^{\infty} \frac{z^2}{1+z^2} dz$$


0.50
0.40
0.30
0.20
0.10
0.00
Lorentzian/Gaussian distributions
μ
Gaussian
σ
Γ
Lorentzian
−5
−4
−3
−2
−1
0
1
2
3
4
5
x

The *F*-Distribution



The *F*-Distribution

The F distribution is the ratio of two variance estimates:

$$F = \frac{s_1^2}{s_2^2} = \frac{est.\sigma_1^2}{est.\sigma_2^2}$$

Also the ratio of two chi-squares, each divided by its degrees of freedom:

$$F = \frac{\chi^2_{(v_1)} / v_1}{\chi^2_{(v_2)}) / v_2}$$

- $v_2 > v_1$, and $v_2 > 2$.

Then the mean of the F distribution (expected value) = $v_2 / (v_2 - 2)$

F-Distribution (pt.2)

F depends on v_1 and v_2 (df_1 and df_2).

These dictate the shape of *F*. Range is 0 to infinity.

F tables show critical values for *df* in the numerator and *df* in the denominator.

F tables are 1-tailed (2-tailed are atypical)

A continuous distribution

ANOVA

ANalysis Of VAriance

Statistical model that estimates variance thorough differences in means

Developed by Sir Ronald Fisher

- British Statistician & Geneticist

Source of Variation	df	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F-test	p-value
Treatment	k-1	SSTr	MSTr=SSTr/(k-1)	F=MSTr/MSE	
Error	N-k	SSE	MSE=SSE/(N-k)		
Total	N-1	SSTo			

https://courses.lumenlearning.com/suny-natural-resources-biometrics/chapter/chapter-5-one-way-analysis-of-variance/

Summary of Major Distributions

1) To understand error you MUST understand the distributions you are dealing with

2) But what about multiple measures:
   - Error Propagation

Experimental Design

- Proper application of biostatistics involves experimental design
- Experimental design is done before testing
- Proper application of experimental design and statistical analysis can save a lot of:
  - Time
  - Resources
  - lives (potentially), etc.

Hypothesis Testing

- test a hypothesized value

- need some value based on past experience, claims of other people, dream your thesis advisor had, etc.

- new situation (e.g. new growth hormone) produces results that are no different, on average, from those results previously occurring.

→ Null Hypothesis

e.g. Weight training, with a healthy diet, produces a mean increase in muscle mass of 12.6kg, over 6 weeks. A new growth hormone, yulegosterol, increases this.

$H_o = \mu = \mu_o = 12.6$ (Null hypothesis, nothing different)

$H_A = \mu > \mu_o = 12.6$ (Alternative, there is weight gain)

Hypothesis Testing cont

- this is a one-sided alternative (and one sided test)

- if we have no clue on what new hormone will really do

We can use a 2 sided alternative

Ho= $\mu = \mu_o = 12.6$ (Null hypothesis, nothing different)

HA = $\mu \neq \mu_o = 12.6$ (Alternative, $\pm$ weight gain)

- Regardless of whichever, we still need some criterion for deciding how far away $\bar{x}$ can be from μ_o before we reject Ho.

- choose a level of significance (e.g. α=0.05). This is a level of probability that the test will fail.

The appropriate Test in this Example: Student's T-test

8 volunteers (degrees of freedom, df = 8-1=7)

mean weight gain = 20.2kg

standard deviation = 4.3

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

From Student's t-test table (8-1 = 7df; α=0.05):

- reject Ho if t > 1.895 (one sided)
- reject Ho if t > +2.365 or t < -2.365 (two sided)

(i.e. reject if | t | > 2.365)

ASSUMES NORMALLY DISTRIBUTED DATA

- This should be tested

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
						Confidence Level					

$t_{.95}$	$t_{.975}$
0.05	**0.025**
0.10	**0.05**
6.314	12.71
2.920	4.303
2.353	3.182
2.132	2.776
2.015	2.571
1.943	2.447
1.895	2.365
1.860	2.306
1.833	2.262
1.812	2.228

Confidence Intervals

- what range of values are we confident that the measurements can take ?
- theory states that 95% of the time for a t value with df = 7:

$$-2.365 \leq t \leq +2.365$$

$$-2.365 \leq \left[\frac{\bar{x} - \mu}{s/\sqrt{n}} \right] \leq +2.365$$

A little algebra…...

$$\boxed{\left(\bar{x} - 2.365 \cdot s/\sqrt{n} \right) \leq \mu \leq \left(\bar{x} + 2.365 \cdot s/\sqrt{n} \right)}$$

Test for Two Means

Calculation of Denominator depends on:

1. The 2 populations having common variance, σ^2
2. The 2 σ^2s, or common σ^2 is known or estimated
3. If both samples are (or are not) the same size
4. Paired vs. Independent

$$t = \frac{\overline{x_1} - \overline{x_2}}{s_{\overline{x_1} - \overline{x_2}}}$$

The Choice of Rejection depends on:

1. The level of significance chosen (α)
2. The sample size (n)
3. The test required (i.e. 1 or 2 tailed)

Independent vs. Paired

Pairing:

- done prior to experiment on the basis of similar responses in the absence of treatment effects.
- e.g. comparing drug therapy in sets of identical twins
- - if members of a pair tend to be positively correlated an increase in the ability of the experiment to detect a small difference is possible.

Independent

- Compares the means between 2 groups

2 Means, Independent Samples, Equal Variances:

$H_0 : \mu_1 = \mu_2$ (Null Hypothesis)

$$t = \frac{\left(\overline{x_1} - \overline{x_2}\right) - \left(\mu_1 - \mu_2\right)}{s_{\overline{x_1} - \overline{x_2}}}$$

Need weighted average of sample variances:

$$s^2 = \frac{\left(n_1 - 1\right)s_1^{\ 2} + \left(n_2 - 1\right)s_2^{\ 2}}{\left(n_1 - 1\right) + \left(n_2 - 1\right)}$$

Note df = $(n_1 - 1) + (n_2 - 1)$

Situation #1: $n_1 \neq n_2$

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{s^2\left(\frac{n_1 + n_2}{n_1 n_2}\right)}$$

Situation #2: $n_1 = n_2$

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s^2}{n}}$$

Comparing Paired Sample Means

- compute differences in pairs
- calculate average pair difference

$$s = \sqrt{\frac{\sum_j D_j^2 - \left(\sum_j D_j\right)^2 \Big/ n}{n-1}}$$

$$t = \frac{\overline{D}}{s/\sqrt{n}}$$

j = number of pairs

Also note here n = j

Note: df = j - 1

Testing the Hypothesis of Equality of Variances (homoscedasticity):

- up to now it is assumed that variances are equal, based on some pre-decided criterion.

How is this tested ?

Null hypothesis: $\sigma_1^2 = \sigma_2^2$

$$F_{\alpha,m,n} = \frac{s^2_{BIG}}{s^2_{SMALL}}$$

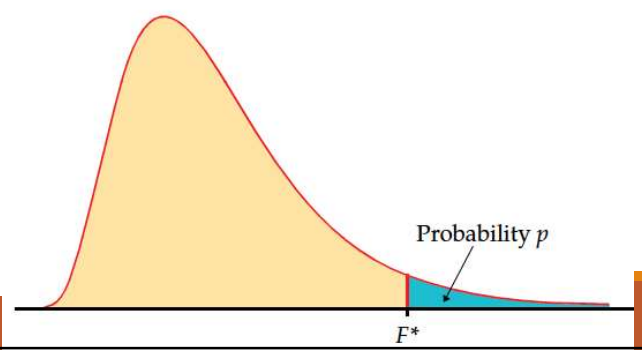
m-1 = *df* for s^2_{big}

n-1 = *df* for s^2_{small}

α = level of significance desired (e.g. 0.05 for 95% confidence)

F-distribution

Degrees of freedom in the denominator	p	Degrees of freedom in the numerator: 1	2	3	4	5	6	7	8	9
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
	.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
	.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
	.001	405284	500000	540379	562500	576405	585937	592873	598144	602284
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39
3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
	.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86
4	.100	4.54	4.32	4.19	4.11					
	.050	7.71	6.94	6.59	6.39					
	.025	12.22	10.65	9.98	9.60					
	.010	21.20	18.00	16.69	15.98					
	.001	74.14	61.25	56.18	53.44					
5	.100	4.06	3.78	3.62	3.52					
	.050	6.61	5.79	5.41	5.19					
	.025	10.01	8.43	7.76	7.39					
	.010	16.26	13.27	12.06	11.39					
	.001	47.18	37.12	33.20	31.09					



If 2 samples have unequal Variances

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

$$t' = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1-\bar{x}_2}}$$

effective $$df = \frac{\left(s_1^{\,2}/n_1 + s_2^{\,2}/n_2\right)}{\left[\left(s_1^{\,2}/n_1\right)^2/(n_1-1)\right] + \left[\left(s_2^{\,2}/n_2\right)^2/(n_2-1)\right]}$$

T.TEST function

Returns the probability that is associated with a Student's t-Test. Use **T.TEST** to determine whether two samples are likely to have come from the same two underlying populations that have the same mean.

Syntax

T.TEST(array1,array2,tails,type)

Argument	Description	Remarks
array1	The first data set.	• None.
array2	The second data set.	• None.
tails	Specifies the number of distribution tails.	• If **tails** = 1, **T.TEST** uses the one-tailed distribution. If **tails** = 2, **T.TEST** uses the two-tailed distribution. • If **tails** is any value other than 1 or 2, this function returns the #NUM! error value. • If this argument is nonnumeric, this function returns the #VALUE! error value. • If this argument contains a decimal value, this function ignores the numbers to the right side of the decimal point.
type	The kind of t-Test to perform.	• If **type** equals 1, **T.TEST** performs a paired test. If **type** equals 2, **T.TEST** performs a two-sample equal variance (homoscedastic) test. If **type** equals 3, **T.TEST** performs a two-sample unequal variance (heteroscedastic) test. • If this argument is nonnumeric, this function returns the #VALUE! error value. • If this argument contains a decimal value, this function ignores the numbers to the right side of the decimal point.

MATLAB

ttest

One-sample and paired-sample *t*-test

expand all in page

Syntax

```
h = ttest(x)                                   example

h = ttest(x,y)                                 example
h = ttest(x,y,Name,Value)                      example

h = ttest(x,m)                                 example
h = ttest(x,m,Name,Value)                      example

[h,p] = ttest( __ )                            example
[h,p,ci,stats] = ttest( __ )                   example
```

Description

`h = ttest(x)` returns a test decision for the null hypothesis that the data in `x` comes from a normal distribution with mean equal to zero and unknown variance, using the one-sample *t*-test. The alternative hypothesis is that the population distribution does not have a mean equal to zero. The result `h` is `1` if the test rejects the null hypothesis at the 5% significance level, and `0` otherwise. example

What about Multiple Comparisons?

- There are more powerful techniques
- t-test is not appropriate
  - get compounding error

e.g.

- Analysis of Variance (ANOVA)
- comparisons of multiple treatments

Power, Sample Size, and the Detection of Differences

The error rate or significance level is chosen = α

(e.g. $\alpha = 0.05$)

TYPE I Error

- α - we make a mistake and falsely reject HO

TYPE II Error

- β - we make a mistake and falsely accept HO

	Decision	Data from a population for which:	
		H_0 is TRUE, H_1 false	H_0 is false, H_1 TRUE
Non-significant	Accept H_0 Reject H_1	*Correct Decision* Probability should be high. Symbol: $1 - \alpha$ = **Confidence coefficient**	*Incorrect Decision* ➔ Type II error made Probability should be low. Symbol: β
Significant	Reject H_0 Accept H_1	*Incorrect Decision* ➔ Type I error made Probability should be low. Symbol: α (significance level)	*Correct Decision* Probability should be low. Symbol: $1 - \beta$ = **power**

Errors

- if we use $\alpha = 10\%$
  - there is a high tendency to conclude that HO will be false when it is not
- if use $\alpha = 0.1\%$
  - then you are unlikely to erroneously state that HO is rejected.
  - Tests such as this are conservative and reliable
  - Also fairly unlikely to state that HO is rejected when in truth it should be accepted

	Value of α	
Type of test:	10% (liberal)	0.1% (cautious, conservative)
If H_0 is true	May well reject H_0	Unlikely to reject H_0
If H_0 is false	Good chance of rejecting H_0	Some chance of rejecting H_0
If H_0 is not rejected	Very little reason found to distrust H_0	Support for H_0 may not be impressive
If H_0 is rejected	Possibly over-hasty rejection of H_0	Very convincing evidence against H_0

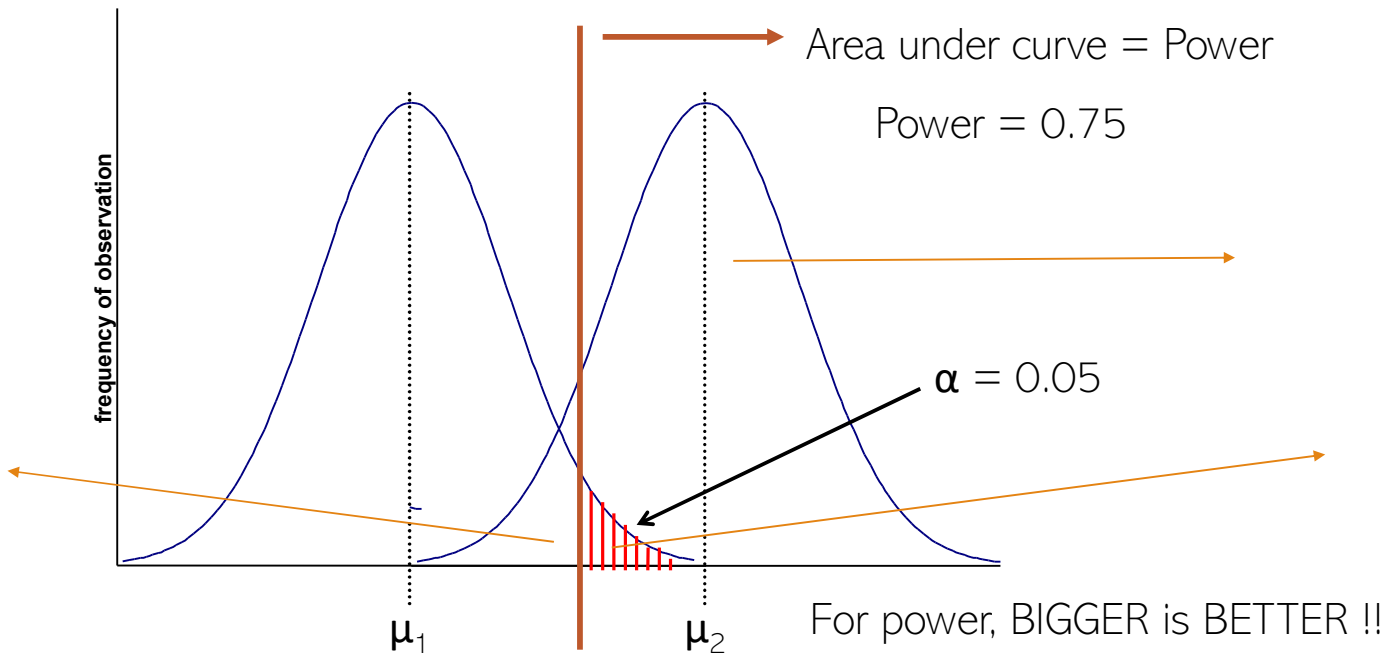
Power

- a test of significance should reject HO when it is really false
- the probability a test does this is the Power

Power is a complex quantity depending on:

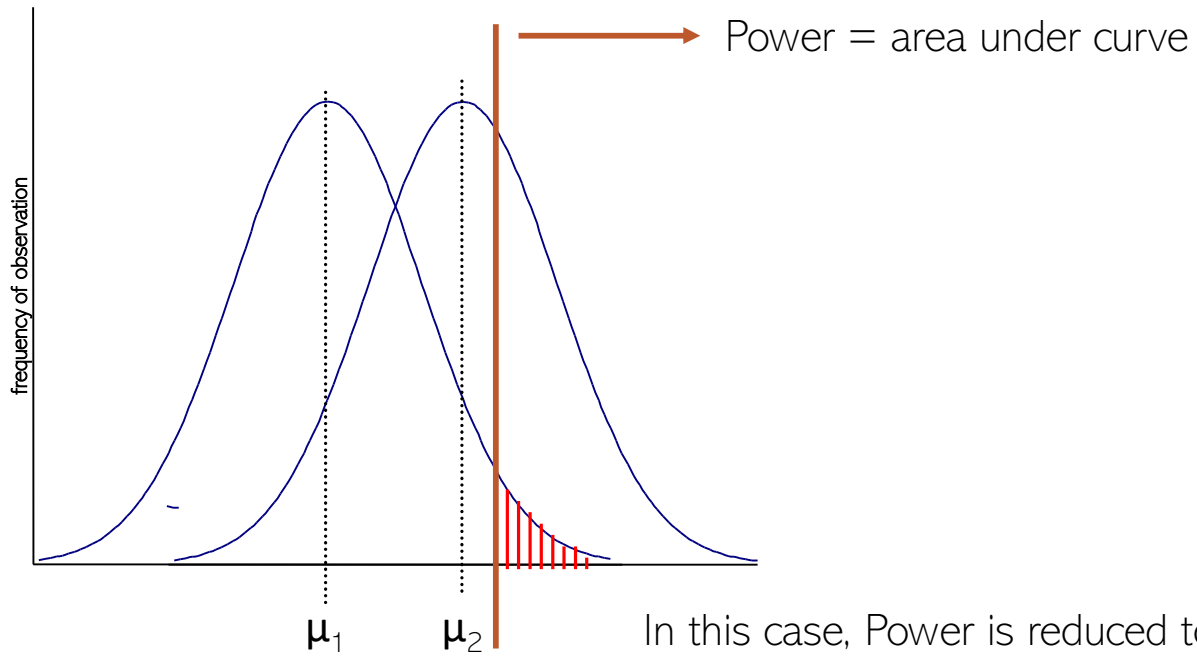
- chosen α
- variance σ
- number in sample, n
- difference in means (i.e. $\mu 1$ - $\mu 2$)

Graphical Representation of Power



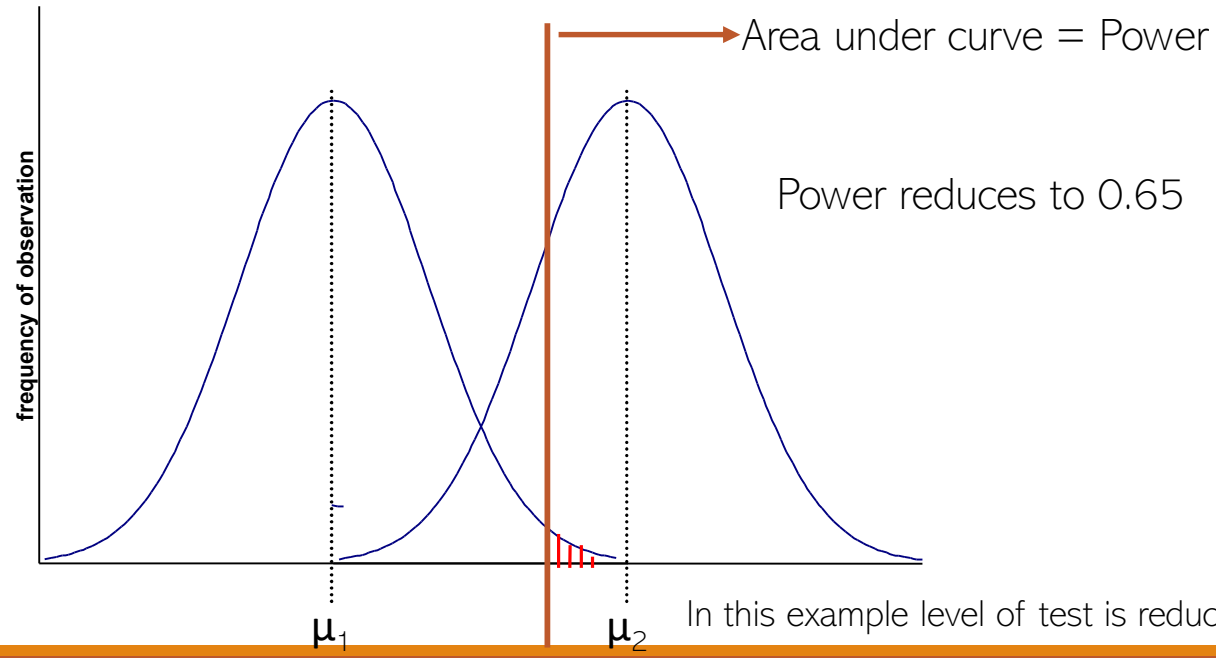
For power, BIGGER is BETTER !!

Graphical Representation of Power



In this case, Power is reduced to 0.455

Graphical Representation of Power



In this example level of test is reduced to $\alpha = 0.01$

- for a 1-tailed test, at level **α**, the power is the probability that the normal deviate

$$Z > \left(-\sqrt{n}/\sigma\right)\left(\mu_1 - \mu_2\right) + Z_{2\alpha}$$

$Z_{2\alpha}$ = 1.645 for a 1-tailed test with **α** = 0.05 (i.e. 5% significance level)
$Z_{2\alpha}$ = 2.326 for a 1-tailed test with **α** = 0.01 (i.e. 1% significance level)

The key factor in which power depends is:

$$\boxed{\phi = \sqrt{n}\left(\mu_1 - \mu_2\right)/\sigma}$$

(Single or paired samples)

Note: Use sqrt(n/2) if samples are independent !

Sample Power Calculations

$$\phi = \sqrt{n}\left(\mu_1 - \mu_2\right)/\sigma =$$

Level of Test	# tails	1.5	2	2.5	3	3.5
0.05	1	0.44	0.64	0.80	0.91	0.97
	2	0.32	0.52	0.71	0.85	0.94
0.01	1	0.20	0.37	0.57	0.75	0.88
	2	0.14	0.22	0.47	0.61	0.82

<u>example:</u> Let's say there are 5 samples already measured (n=10) where the difference between means is, **Δμ**=18.2-16.6=1.6, and the standard deviation is **σ**=2.6. What is the power ?

So, how many samples (n) are needed anyway?

One Quick Method…..

1) decide on the approximate desired power wanted for a

specific value of μ1 - μ2

2) use table (previous slide) to determine the approximate value needed of ϕ for the intended level of significance and nature(i.e. 1 or 2 tailed) of the test.

3) Use formula to solve for n:

$$n = \left[\frac{\phi\sigma}{(\mu_1 - \mu_2)}\right]^2$$

Paired, or single samples

$$n = 2 \cdot \left[\frac{\phi\sigma}{(\mu_1 - \mu_2)}\right]^2$$

Independent samples

Type I Error Rate and Multiple T-tests

- consider no [true] difference between 2 populations

- by random chance alone there is a 100*α% chance of declaring an [incorrect] difference between the two populations.

- error compounded when multiple t-tests are carried out.

if k independent t-tests are performed with α level of significance, then the probability of observing no significant (X) differences is:

Type I Error Rate and Multiple T-tests

- the probability of observing at least one significant difference (when none exist) is:

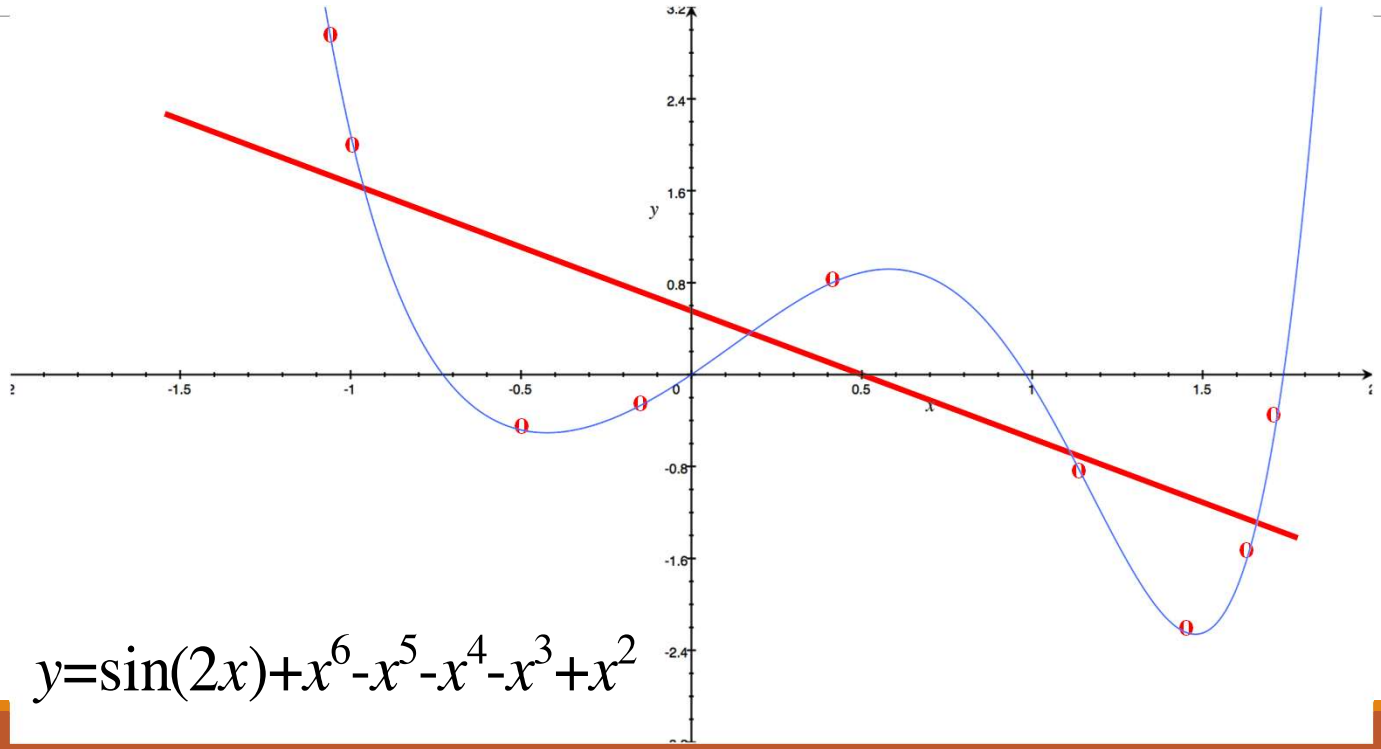
Thus as k ↑ the probability of a Type I error ↑.

e.g. if 10 independent t- tests are carried out, the probability of declaring at least one significant difference (even though there are none) is:

Table 1: Type I error rate for k independent tests with a significance level of Alpha

	Alpha			
k	0.1	0.05	0.01	0.001
1	0.1	0.05	0.01	0.001
2	0.19	0.098	0.02	0.002
3	0.271	0.143	0.03	0.003
4	0.344	0.185	0.039	0.004
5	0.41	0.226	0.049	0.005
6	0.469	0.265	0.059	0.006
7	0.522	0.302	0.068	0.007
8	0.57	0.337	0.077	0.008
9	0.613	0.37	0.086	0.009
10	0.651	0.401	0.096	0.01

What is the Appropriate Mathematical Model: How to choose?



$$y=\sin(2x)+x^6-x^5-x^4-x^3+x^2$$

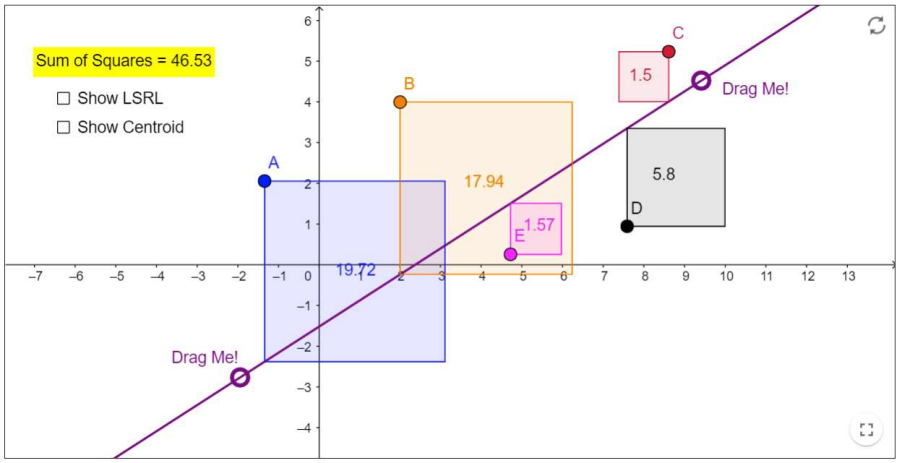
Least Squares Regression

Statistical method to show a relationship between x and y variables

The least squares regression function makes the vertical distance from the data points to the regression line the smallest

Minimizes the variance (sum of squares error)

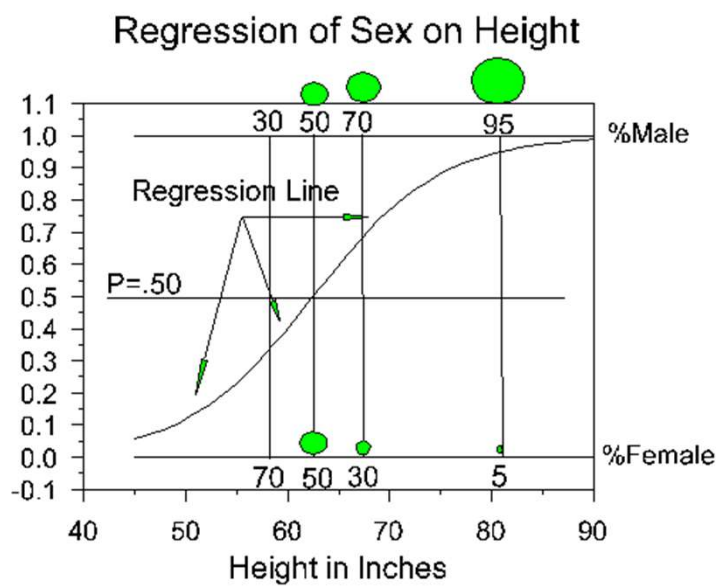
"Regression" generally refers to simple linear least squares regression



Example

Logistic Regression

- Independent variable (x) vs Nominal dependent variable (y)
- Regression line is still average but nonlinear
- Data points don't fall on regression line
- Example:
  - T –test look at Null hypothesis that cell reproduction rate is not linked to a tissue being cancerous
  - Logistic regression – predict the probability that tissue with a specific cellular reproduction rate will end up metastasizing in the next 5 months



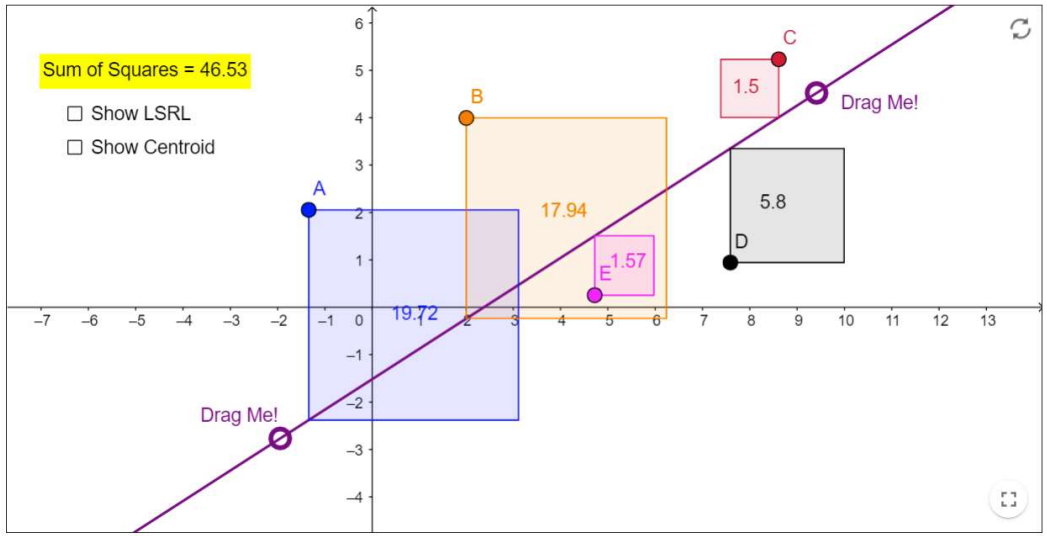
http://faculty.cas.usf.edu/mbrannick/regression/Logistic.html

Linear Regression

Statistical method to show a **linear** relationship between x and y variables

Finds the **line** of best fit (regression line)

The Least Squares Regression Line makes the vertical distance from the data points to the regression line the smallest



Linear Regression

The equation for a straight line:

$$Y = \beta 1X + \beta 0$$

$\beta 0$ = intercept
$\beta 1$ = slope

For a data point, I (x_i, y_i), scattered about the line, its position can be represented by:

$$Y_i = \beta 1X_i + \beta 0 + \varepsilon_i$$

where ε represents the deviation of the point from the line

Therefore, our set of observations can be represented by a set of equations:

$$Y1 = \beta 1X1 + \beta 0 + \varepsilon 1$$
$$Y2 = \beta 1X2 + \beta 0 + \varepsilon 2$$
$$Y3 = \beta 1X3 + \beta 0 + \varepsilon 3$$
$$Y4 = \beta 1X4 + \beta 0 + \varepsilon 4$$

.....

$$Yn = \beta 1Xn + \beta 0 + \varepsilon n$$

Linear Regression

This set of equations can then be written in the form of the vectors and matrices:

$$Y = X\beta + \varepsilon$$

$$Y = \begin{vmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{vmatrix} \quad X = \begin{vmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{vmatrix} \quad \beta = \begin{vmatrix} \beta_0 \\ \beta_1 \end{vmatrix} \quad \varepsilon = \begin{vmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{vmatrix}$$

$$\boxed{\begin{vmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{vmatrix} = \begin{vmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{vmatrix} \cdot \begin{vmatrix} \beta_0 \\ \beta_1 \end{vmatrix} + \begin{vmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{vmatrix}}$$

$$Y_{i,\,ave} = X_i\,\beta \qquad \boldsymbol{\varepsilon} = Y_i - Y_{i,\,ave}$$

Linear Regression

The overall failure of the data to fit the model is the residual, ε, sum of squares, $\Sigma\ \varepsilon_i^2$:

$$|\varepsilon_1\ \varepsilon_2\ \cdots\ \varepsilon_n| \cdot \begin{vmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{vmatrix} = \sum \varepsilon^2$$

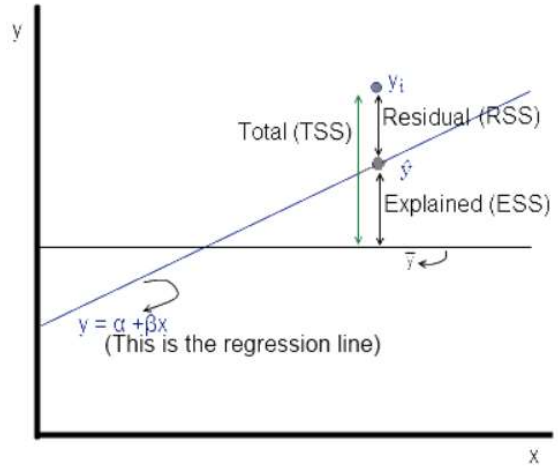
According to the method of least mean squares, we want to minimize $\Sigma\ \varepsilon_i^2$. The equation that provides this estimates of b_0 and b_1 is:

We are interested in $\boldsymbol{\beta}$:

where: $$(X'X)^{-1} = \frac{1}{n\sum X_i^2 - (\sum X_i)^2} \begin{pmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{pmatrix}$$

From this set of equations you can derive slope and intercept. After calculating sums you just need to perform simple matrix algebra to determine the intercept, β_0, and slope, β_1.

ANOVA



$\hat{y}$ is the predicted value of y given x, using the equation $\hat{y} = \alpha + \beta x$.

y_i is the actual observed value of y

$\bar{y}$ is the mean of y.

The distances that RSS, ESS and TSS represent are shown in the diagram to the left - but remember that the actual calculations are squares of these distances.

$$TSS = \Sigma (y_i - \bar{y})^2$$

$$RSS = \Sigma (y_i - \hat{y})^2$$

$$ESS = \Sigma (\hat{y} - \bar{y})^2$$

https://www.riskprep.com/component/exam/?view=exam&layout=detail&id=131

Source of Variation	df	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F-test	p-value
Treatment	k-1	SSTr	MSTr=SSTr/(k-1)	F=MSTr/MSE	
Error	N-k	SSE	MSE=SSE/(N-k)		
Total	N-1	SSTo			

https://courses.lumenlearning.com/suny-natural-resources-biometrics/chapter/chapter-5-one-way-analysis-of-variance/

ANOVA

→ Is this the appropriate model?

Source	SS *(Sum of Squares, the numerator of the variance)*	DF *(the denominator)*	MS *(Mean Square, the variance)*	F
Regression *(or Model)*	$SSR = \sum_{i=1}^{n} ((\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y})^2$	*2-1=1*	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Error	$SSE = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$	*n-2*	$MSE = \frac{SSE}{n-2}$	
Total	$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$	*n-1*		

df = "degrees of freedom"

F = Calculated F. This is compared to $F_{\alpha,df(R),df(E)}$

Multiple Regression

- Relationship between a dependent variable, Y, and several independent variables which simultaneously influence the dependent variable.
- **β**'s are called the **partial regression coefficients**
  - β_1 represents the true change in the mean of Y when X_1 changes by 1 unit, and all other variables are held constant
  - similarly for β_2, β_3, and $\beta_{4\ldots\ ETC}$.

Multiple Regression Example

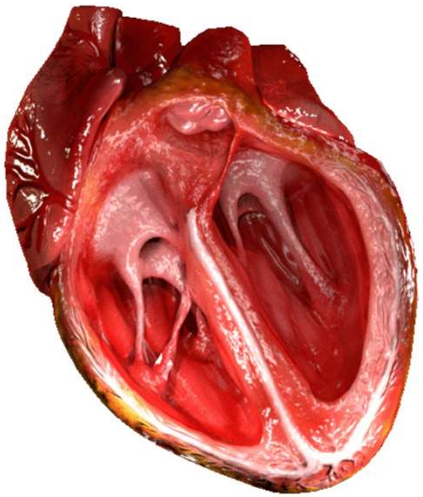
What is the relationship between left myocardial contractile force and serum ionic composition ?

response (dependent variable):

- contractile force (Y)

independent variables

- Chloride (Cl) (X1)
- phosphate (PO_4) (X2)
- Potassium (K) (X3)
- Sodium (Na) (X4)



The Data:

Y = contractile force (N)
X_1 = chloride (units)
X_2 = phosphate (units)
X_3 = potassium (units)
X_4 = sodium (units)

- measured in isolated heart preparations.

	X1 (Cl)	**X2** (PO4)	**X3** (K)	**X4** (Na)	**Y** (force)
1	2.2	0.417	1.35	1.79	351
2	2.1	0.354	0.9	1.08	249
3	1.52	0.208	0.71	0.47	171
4	2.88	0.335	0.9	1.48	373
5	2.18	0.314	1.26	1.09	321
6	1.87	0.271	1.15	0.99	191
7	1.52	0.164	0.83	0.85	225
8	2.37	0.302	0.89	0.94	291
9	2.06	0.373	0.79	0.8	284
10	1.84	0.265	0.72	0.77	213
11	1.89	0.192	0.46	0.46	138
12	2.45	0.221	0.76	0.95	213
13	1.88	0.186	0.52	0.95	151
14	1.93	0.207	0.6	0.92	130
15	1.8	0.157	0.67	0.6	93
16	1.81	0.195	0.47	0.57	95
17	1.49	0.165	0.66	0.8	147
18	1.53	0.226	0.68	0.66	88
19	1.43	0.224	0.44	0.45	65
20	1.54	0.271	0.51	0.95	120
21	1.13	0.187	0.38	0.63	72
22	1.63	0.2	0.62	1.1	160
23	1.36	0.211	0.71	0.47	72
24	1.76	0.283	0.96	0.96	252
25	2.53	0.284	0.85	1.39	310
26	2.59	0.303	1.02	0.95	336
TOTALS	49.29	6.515	19.81	23.07	5111

Example: Regression Equation

In the absence of any biological understanding of how contractile force may be related to blood ion composition, it's best to first choose a linear model:

for i = 1,…,26

<u>Assumptions</u>:

- $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_{26}$ are a random sample from a normal population with a mean = 0 and some constant (unknown) $\boldsymbol{\sigma}^2$. [i.e. $\boldsymbol{\varepsilon}_i \sim N(0, \boldsymbol{\sigma}^2)$]

- The relationship between Y and each X (all other X's held constant) is linear.

- The X's do not interact

Example: Estimating the Parameters

- use least squares analysis.

i.e. choose estimates of β as [b0, b1, b2, b3, b4] which minimize:

$$\sum_{i=1}^{n}(Y_i - \hat{\mu}_{Y.1234})^2 = \sum_{i=1}^{n}(Y_i - b_0 - b_1X_{i1} - b_2X_{i2} - b_3X_{i3} - b_4X_{i4})^2$$

(equation 1)

- this is a calculus problem. Differentiating the above expression, with respect to each b, results in 5 equations in the 5 variables (b0, b1, b2, b3, b4) to be solved.

Example: Normal Equations

$$b_0 n + b_1 X_{\bullet 1} + b_2 X_{\bullet 2} + b_3 X_{\bullet 3} + b_4 X_{\bullet 4} = Y_{\bullet} = \sum Y$$

$$b_0 X_{\bullet 1} + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 + b_3 \sum X_1 X_3 + b_4 \sum X_1 X_4 = \sum X_1 Y$$

$$b_0 X_{\bullet 2} + b_1 \sum X_2 X_1 + b_2 \sum X_2^2 + b_3 \sum X_2 X_3 + b_4 \sum X_2 X_4 = \sum X_2 Y$$

$$b_0 X_{\bullet 3} + b_1 \sum X_3 X_1 + b_2 \sum X_3 X_2 + b_3 \sum X_3^2 + b_4 \sum X_3 X_4 = \sum X_3 Y$$

$$b_0 X_{\bullet 4} + b_1 \sum X_4 X_1 + b_2 \sum X_4 X_2 + b_4 \sum X_4 X_3 + b_4 \sum X_4^2 = \sum X_4 Y$$

(equations 2)

- the solutions to these equations is only simple high school algebra.

Example: Matrix Representation

Equations can be rewritten using matrix notation.

Model:

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} + \beta_4 X_{14} + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \beta_3 X_{23} + \beta_4 X_{24} + \varepsilon_2$$

$$\vdots$$

$$Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \beta_3 X_{n3} + \beta_4 X_{n4} + \varepsilon_n$$

Matrix Notation:

$$\boxed{Y = X\underline{\beta} + \underline{\varepsilon}}$$

(equation 3)

Example: Matrix Representation

where:

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} & X_{13} & X_{14} \\ 1 & X_{21} & X_{22} & X_{23} & X_{24} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & X_{n3} & X_{n4} \end{bmatrix} = \begin{bmatrix} 1 & 2.20 & 0.417 & 1.35 & 1.79 \\ 1 & 2.10 & 0.354 & 0.90 & 1.08 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2.59 & 0.303 & 1.02 & 0.95 \end{bmatrix}_{26\times 5}$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 351 \\ 249 \\ \vdots \\ 336 \end{bmatrix}_{26\times 1} \qquad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}_{5\times 1} \qquad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{26} \end{bmatrix}_{26\times 1}$$

Example: Matrix Representation

Furthermore, the Normal Equations (equations 2) can be rewritten as the Matrix Equation…

$$\boxed{(X'X)\cdot b = X'Y}$$ *(equation 4)*

NOTE: X' X is a symmetric matrix (about the diagonal)

$$X'X = \begin{bmatrix} n & \sum X_1 & \sum X_2 & \sum X_3 & \sum X_4 \\ \sum X_1 & \sum X_1^2 & \sum X_1X_2 & \sum X_1X_3 & \sum X_1X_4 \\ \sum X_2 & \sum X_2X_1 & \sum X_2^2 & \sum X_2X_3 & \sum X_2X_4 \\ \sum X_3 & \sum X_3X_1 & \sum X_3X_2 & \sum X_3^2 & \sum X_3X_4 \\ \sum X_4 & \sum X_4X_1 & \sum X_4X_2 & \sum X_4X_3 & \sum X_4^2 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} \sum Y \\ \sum X_1Y \\ \sum X_2Y \\ \sum X_3Y \\ \sum X_4Y \end{bmatrix}$$

$$b_0 n + b_1 X_{\bullet 1} + b_2 X_{\bullet 2} + b_3 X_{\bullet 3} + b_4 X_{\bullet 4} = Y_{\bullet} = \sum Y$$

$$b_0 X_{\bullet 1} + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 + b_3 \sum X_1 X_3 + b_4 \sum X_1 X_4 = \sum X_1 Y$$

$$b_0 X_{\bullet 2} + b_1 \sum X_2 X_1 + b_2 \sum X_2^2 + b_3 \sum X_2 X_3 + b_4 \sum X_2 X_4 = \sum X_2 Y$$

$$b_0 X_{\bullet 3} + b_1 \sum X_3 X_1 + b_2 \sum X_3 X_2 + b_3 \sum X_3^2 + b_4 \sum X_3 X_4 = \sum X_3 Y$$

$$b_0 X_{\bullet 4} + b_1 \sum X_4 X_1 + b_2 \sum X_4 X_2 + b_4 \sum X_4 X_3 + b_4 \sum X_4^2 = \sum X_4 Y$$

$$X'X = \begin{bmatrix} n & \sum X_1 & \sum X_2 & \sum X_3 & \sum X_4 \\ \sum X_1 & \sum X_1^2 & \sum X_1 X_2 & \sum X_1 X_3 & \sum X_1 X_4 \\ \sum X_2 & \sum X_2 X_1 & \sum X_2^2 & \sum X_2 X_3 & \sum X_2 X_4 \\ \sum X_3 & \sum X_3 X_1 & \sum X_3 X_2 & \sum X_3^2 & \sum X_3 X_4 \\ \sum X_4 & \sum X_4 X_1 & \sum X_4 X_2 & \sum X_4 X_3 & \sum X_4^2 \end{bmatrix} \qquad X'Y = \begin{bmatrix} \sum Y \\ \sum X_1 Y \\ \sum X_2 Y \\ \sum X_3 Y \\ \sum X_4 Y \end{bmatrix}$$

(equations 2)

Example: Solve for b

The solution to Equation 4 can also be rewritten as:

$$\underline{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} = (X'X)^{-1} \cdot X'Y$$

where $(X'X)^{-1}$ denotes the inverse of the matrix $X'X$. This now looks simple but it still requires as much computation as the normal equations.

- The advantage to the matrix approach is in book keeping (neater to write down equations and keep track of data).

Example: Sub in real numbers

$$X'X = \begin{bmatrix} 26.0 & 49.29 & 6.515 & 19.81 & 23.07 \\ 49.29 & 97.9781 & 12.7981 & 39.0012 & 45.9843 \\ 6.515 & 12.7981 & 1.7540 & 5.2688 & 6.1600 \\ 19.81 & 39.0012 & 5.2688 & 16.6387 & 18.9306 \\ 23.07 & 45.9843 & 6.1600 & 18.9306 & 23.1015 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 5111.0 \\ 10521.2 \\ 1409.48 \\ 4363.95 \\ 5129.58 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 0.9186 & -0.3780 & -0.9663 & -0.1060 & 0.1796 \\ -0.3780 & 0.4209 & -0.6846 & -0.0550 & -0.2328 \\ -0.9663 & -0.6846 & 20.1257 & -2.3701 & -1.0967 \\ -0.1060 & -0.0550 & -2.3701 & 1.5028 & -0.3842 \\ 0.1796 & -0.2328 & -1.0967 & -0.3842 & 0.9346 \end{bmatrix}$$

Example: Solution

$$b = \left(X'X\right)^{-1} \cdot X'Y$$

Therefore,

$$b = \begin{bmatrix} 0.9186 & -0.3780 & -0.9663 & -0.1060 & 0.1796 \\ -0.3780 & 0.4209 & -0.6846 & -0.0550 & -0.2328 \\ -0.9663 & -0.6846 & 20.1257 & -2.3701 & -1.0967 \\ -0.1060 & -0.0550 & -2.3701 & 1.5028 & -0.3842 \\ 0.1796 & -0.2328 & -1.0967 & -0.3842 & 0.9346 \end{bmatrix} \cdot \begin{bmatrix} 5111.0 \\ 10521.2 \\ 1409.48 \\ 4363.95 \\ 5129.58 \end{bmatrix} = \begin{bmatrix} -185.33 \\ 97.76 \\ 256.97 \\ 126.57 \\ 40.28 \end{bmatrix}$$

Example: Solution

$$\widehat{\mu}_{Y.1234} = Y = -185.33 + 97.8X_1 + 257X_2 + 126.6X_3 + 40.3X_4$$

Interpretation:

- it is estimated that hearts from the population sampled with 1 extra unit of blood chloride will beat with 97.8N of force, if all other components of the blood were held constant.

(similarly for PO4, K, and Na)

- obviously the assumption that linearity is valid will only be true for a certain range (i.e. 50 units of Cl would not result in 50 x 97.8N of extra force as this would not be physiologically possible)

Multiple Regression ANOVA

$$\underbrace{\sum\left(Y_i-\overline{Y}\right)^2}_{\text{(total SS)}}=\underbrace{\sum\left(\widehat{\mu}_{Y.1234}-\overline{Y}\right)^2}_{\text{(Model SS)}}+\underbrace{\sum\left(Y_i-\widehat{\mu}_{Y.1234}\right)^2}_{\text{(Residual SS)}}$$

$$\boxed{Model(SS)=\sum\left(\widehat{\mu}_{Y.1234}-\overline{Y}\right)^2=b'\ \cdot\left(X'\ \ Y\right)-\frac{1}{n}\left(\sum Y_i\right)^2}$$

$$\boxed{Total(SS)=\sum\left(Y_i-\overline{Y}\right)^2=\sum Y_i^2-\frac{1}{n}\left(\sum Y_i\right)^2}$$

$$\boxed{residual(SS)=total(SS)-Model(SS)=\sum Y_i^2-b'\cdot\left(X'Y\right)}$$

ANOVA for the Example

Back to the cardiac force example….

$$Total(SS) = 1232659 - \frac{(5111)^2}{26} = 227954.35$$

$$Model(SS) = \begin{bmatrix} -185.33 & 97.76 & 256.97 & 126.57 & 40.28 \end{bmatrix} \cdot \begin{bmatrix} 5111.0 \\ 10521.2 \\ 1409.48 \\ 4363.95 \\ 5129.58 \end{bmatrix} - \frac{(5111)^2}{26} = 197832.43$$

Null Hypothesis: None of the independent variables are of any value in explaining the variation in myocardial contraction force .

i.e. $$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

Alternative, H_1 : not all β's are zero.

Source	df	SS	MS	F_c	$F_{i,j,\alpha}$
regression	4	197382.43	49458.11	34.48	2.84
residual	21	30121.92	1434.38	-------	-----
TOTAL	25	227954.35	---------	-------	-----

$F_C > F_{i,j,\alpha}$ Therefore, reject H_0; at least one β is not zero.

Model Validation: Assumptions

Assumptions that were made:

1) The true mean of Y has been correctly specified.
2) The var(εi) = σ^2 are constant
3) The ε are independent (i.e. uncorrelated) with one another
4) The ε come from a normal distribution.

It is safe to say that not all of these assumptions will be completely met.

All we really require is that they are approximately true.

$$\mu_Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Assessing Assumptions

There are 2 general ways of assessing whether assumptions are at least approximately true.

1) Overfit the model
   - add additional parameters and retest.
   - maybe there are non-linear terms, or interactions between some of the X's.
   - If it is suspected that $\boldsymbol{\varepsilon}$ are not independent then additional terms called variance components can be added which allows you to see model correlations.

2) Examine residuals
   - these are roughly "estimated" errors and hence reflect the properties of the true errors.
   - Graphical analysis is most commonly performed for residual analysis.

Type of Residual Plots

1) Plot residuals vs. $\hat{\mu}_Y$

2) Plot residuals vs. each independent variable, Xi

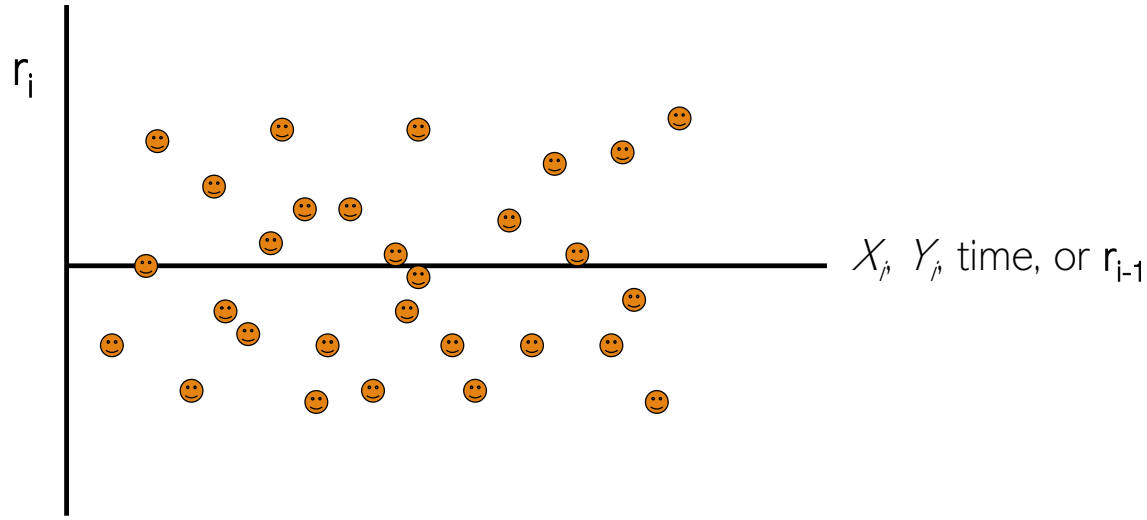
3) Plot residuals vs. time of observation (if appropriate)

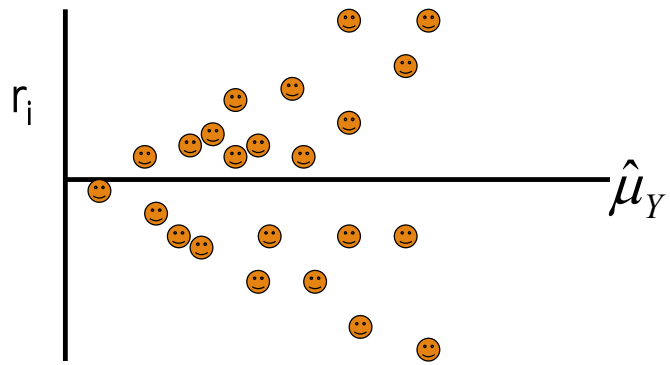
4) Plot ri vs. ri-1 (where i = 2,…,n), to detect serial correlation, assuming the observations are ordered in time or space.

$$r_i = Y_i - \hat{\mu}_Y \text{ (residual = observed - expected)}$$

Residual Plots

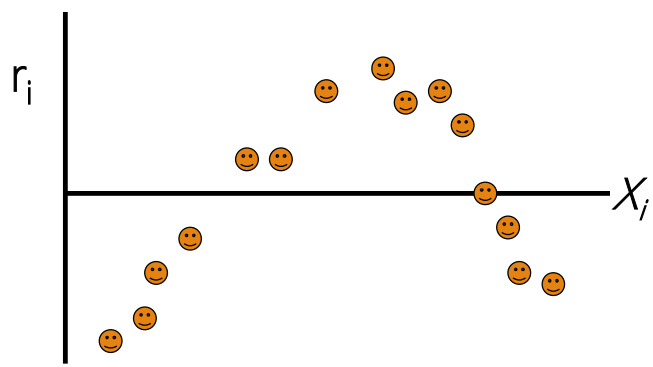
re: residuals should not exhibit any pattern and fall roughly in a band of constant width parallel to the x-axis.





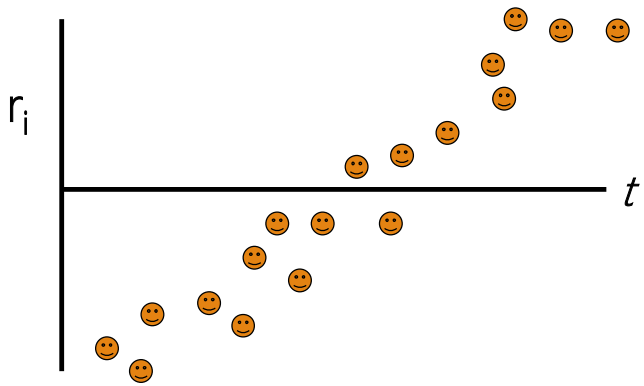
- Assesses assumption of homogeneity of error variance.
- Here error variance ↑ with mean

<u>FIX</u>: Transform Y



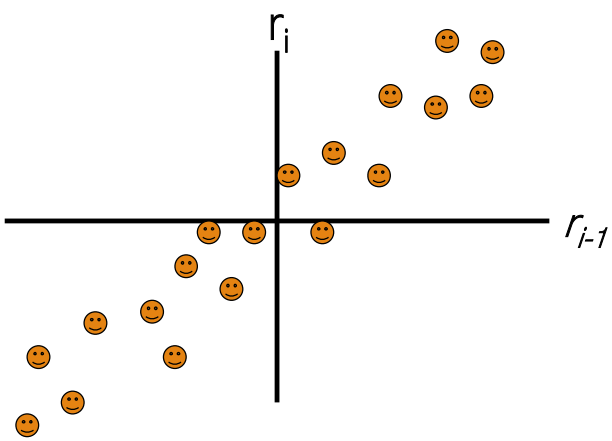
- Detect curvature of relationship between response and X_i

FIX: addition of X_i^2, X_i^3, $\log(X_i)$, or even X_iX_j interaction term(s).



- Response may ↑ or ↓ with time of collection.

FIX: Add time as an independent variable in the model.



- Detects departure from the assumption of independence of errors.
- When data are collected over time, if the errors are not independent they will have a +ve serial correlation.

FIX: modeled variance components.

Outlier

A data point that differs significantly from others within the data set

I.e. lies an abnormal distance from other values in the population

This could be due to experimental error

Could also be due to variability in the measurement

Are generally removed or excluded from data set

Outlier Detection

1) Most algorithms are based on Normal Distributions

2) Typically work one point at a time. But, if more than one point is suspected use multiple outlier test.

3) Approach doesn't work when <6 points to assess

4) Can use mathematical approaches or graphical (e.g. box plot, histogram)

Masking vs Swamping

- a difficult problem!

Masking = too few outliers suggested in the test.

Swamping = specify too many outliers in the test.

One should always complement formal outlier tests with graphical methods.

Swamping and masking are why many tests require that the exact number of outliers being tested is specified

Z-Score and Modified Z-Score

$$Z = \frac{Y_i - \overline{Y}}{s}$$

Ybar = sample mean
s = sample standard deviation

$$M_i = \frac{0.6745(Y_i - \tilde{Y})}{median(|Y_i - \tilde{Y}|)}$$

$\tilde{Y}$ = sample median
denominator = MAD (median absolute deviation)

If $M_i > 3.5$ then there's good chance that value is an outler

Outlier tests

Sample formal outlier tests are grouped by the following characteristics:

1) How are the data distributed? Most tests assume approximately normal distribution.

2) Is the test designed for a single outlier or multiple outliers?

3) If designed for multiple outliers, does the number need to be known exactly or can a range be given?

Outlier Tests

1) Grubbs' Test. Recommended test when testing for a single outlier.

2) Tietjen-Moore Test. This is a generalization of Grubbs' test to account for more than one outlier. It has the limitation that the number of outliers must be specified exactly.

3) Generalized Extreme Studentized Deviate (ESD) Test. Only an upper bound on the suspected number of outliers is needed. Recommended test when the exact number of outliers is not known.

What if want to find the outlier?

Maybe you want to find the outlier (i.e. it is the needle in the haystack you are looking for)

Things get way more complicated!

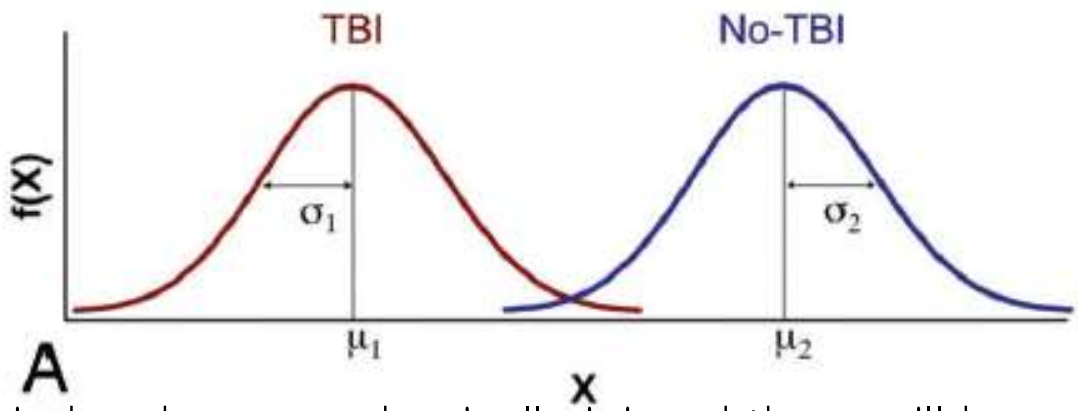
- Anomaly detection
- Data mining

e.g. Use the "Mahalanobis distance" to find outliers.

- Looks at difference between point and a distribution

However this in itself is being effected by the outliers

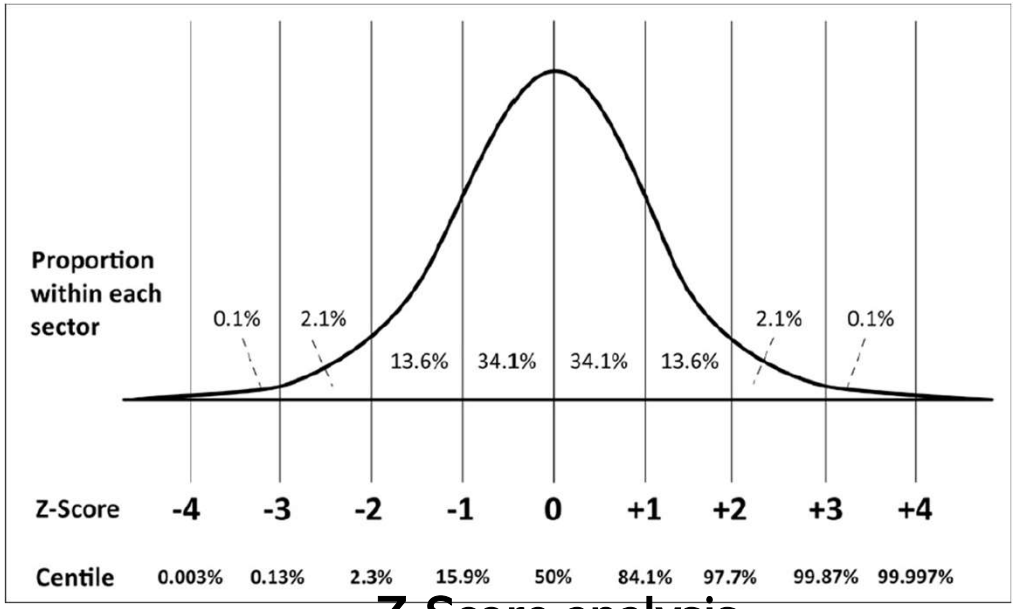
Example



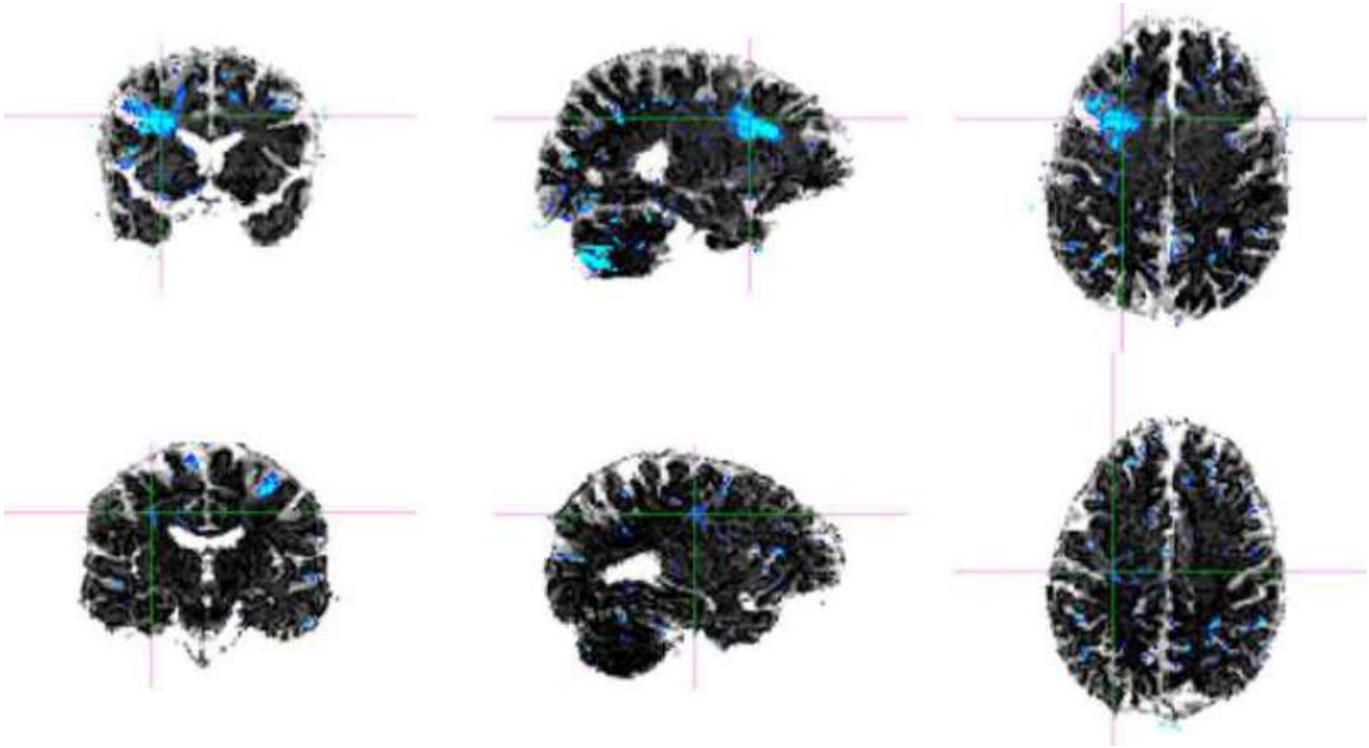
<u>HYPOTHESIS:</u> If the brain has been mechanically injured there will be increase in water diffusivity.

→ Needs to be assessed voxel-by-voxel
→ Requires normative data (~50 controls)
→ All brains need to be spatially warped to a standard template
→ Verification of normality (Skewness and Kurtosis) is critical

Group Analysis



Z-Score analysis



Z-score maps for two patients with chronic mTBI (blue indicates statistically significant areas of free-water compared with the normative atlas)

Brain Imaging and Behavior (2012) 6:137–192

Case-based z-scoring

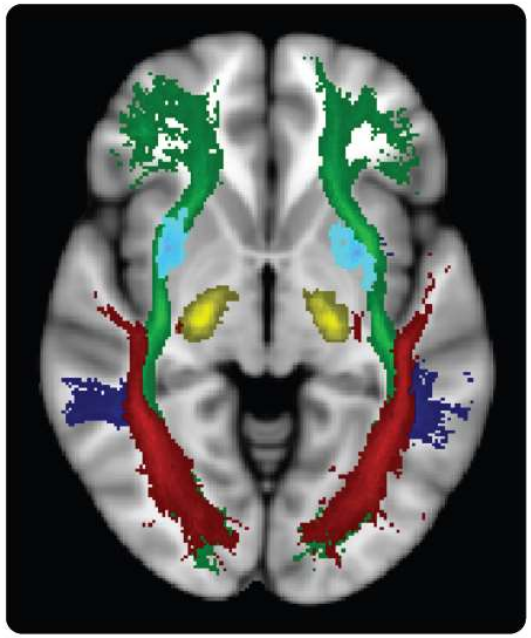
- Identified abnormalities using z-scoring for 24 unique brain ROI within each concussed subject:

$$z = \frac{x - \mu}{\sigma}$$

- ROI is considered abnormal if it is greater than ±2σ relative to the control-group mean

Metric:	Abnormal Region:
FA	Superior longitudinal fasciculus
FA	Inferior longitudinal fasciculus
FA	Inferior fronto-occipital fasciculus
AD, RD	Corticospinal tract
RD	Uncinate fasciculus

*Regions found to be abnormal in >40% of the concussed participants



Outliers, following removal of non-normative control voxels (2-3%), suggestive of damage

ROI	Control Mean	Control SD	No. Outliers (± 2σ)	No. Outliers (± 3σ)
Acoustic Radiation Left	0.2844	0.0257	5	0
Acoustic Radiation Right	0.2724	0.0257	4	0
Cingulate Gyrus Left	0.3609	0.0379	2	0
Cingulate Gyrus Right	0.3095	0.0485	1	0
Cingulum Left	0.3455	0.0407	3	0
Cingulum Right	0.3580	0.0513	0	0
Corpus Callosum	0.4076	0.0319	0	0
Corticospinal Tract Left	0.4713	0.0229	3	2
Corticospinal Tract Right	0.4652	0.0219	5	1
Forceps Major	0.3871	0.0512	0	0
Forceps Minor	0.3770	0.0265	3	1
Fornix	0.2999	0.0415	2	0
Hippocampus Left	0.2652	0.0405	2	0
Hippocampus Right	0.2764	0.0399	1	0
Inferior Fronto-occipital Fasciculus Left	0.3978	0.0207	12	2
Inferior Fronto-occipital Fasciculus Right	0.3945	0.0255	5	0
Inferior Longitudinal Fasciculus Left	0.3517	0.0230	12	1
Inferior Longitudinal Fasciculus Right	0.3600	0.0282	6	1
Optic Radiation Left	0.2975	0.0157	1	0
Optic Radiation Right	0.3092	0.0192	2	0
Superior Longitudinal Fasciculus Left	0.3337	0.0178	11	4
Superior Longitudinal Fasciculus Right	0.3482	0.0185	17	8
Uncinate Fasciculus Left	0.3786	0.0322	7	1
Uncinate Fasciculus Right	0.3534	0.0401	4	0