What is a Chi Square Test?

There are **two types of chi-square tests**. Both use the chi-square statistic and distribution for different purposes:

- A **chi-square goodness of fit test** determines if a sample data matches a population. For more details on this type, see: *Goodness of Fit Test*.
- A **chi-square test for independence** compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of categorical variables differ from each another.
    - A **very small chi square test statistic** means that your observed data fits your expected data extremely well. In other words, there is a relationship.
    - A **very large chi square test statistic** means that the data does not fit very well. In other words, there isn't a relationship.

A chi-square statistic is one way to show a relationship between two categorical variables. In statistics, there are two types of variables: numerical (countable) variables and non-numerical (categorical) variables. **The chi-squared statistic is a single number that tells you how much difference exists between your observed counts and the counts you would expect if there were no relationship at all in the population.**

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

A low value **for chi-square means there is a high correlation between your two sets of data.**

Uses

The chi-squared distribution has many uses in statistics, including:

- Confidence interval estimation for a population standard deviation of a normal distribution from a sample standard deviation.
- Independence of two criteria of classification of qualitative variables.
- Relationships between categorical variables (contingency tables).
- Sample variance study when the underlying distribution is normal.
- Tests of deviations of differences between expected and observed frequencies (one-way tables).
- The chi-square test (a goodness of fit test).

Fstat :-variance among groups

F Value in Regression

The F value in regression is the result of a test where the null hypothesis is that all of the regression coefficients are equal to zero. In other words, the model has no predictive capability. Basically, the f-test compares your model with zero predictor variables (the intercept only model), and decides whether your added coefficients improved the model. If you get a significant result, then whatever coefficients you included in your model improved the model's fit.

The F value in one way ANOVA is a tool to help you answer the question "Is the variance between the means of two populations significantly different?" The F value in the ANOVA test also determines the P value; The P value is the probability of getting a result at least as extreme as the one that was actually observed, given that the null hypothesis is true.

- **MEAN is more influenced by Oulier**
- **Median is not influenced by Outlier**

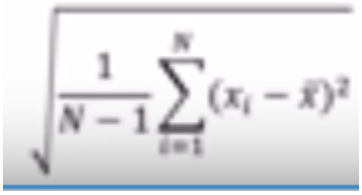
MEASURES OF CENTRAL TENDENCY:-

Mean /Median /Mode

MEASURES OF DISPERSION

Range

SD



Why do we use Squares??

There will be deviations which will be + and some with –ve. So it might tend to 0 which will not capture the variance we want to calculate. So we use square.

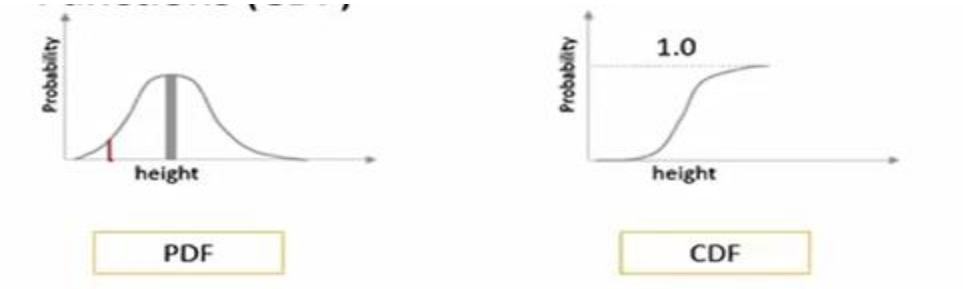
Why do we use SD and not Variance

Variance will not be expressive in units as it is unit square as well

N-1 ➔ degree of freedom. We can move rest variables freely and still maintain the final statistic

Implications are that the effect of square

IQR 75^{th} percentile – 25 percentile



Probability Density functions (PDFs) and Cumulative Density Functions (CDF)

BErnouli Dist:-

One trial and probability of first success /failure

CDF is cumulative

- **Binomial**
  - What is it + Example: Toy problem
  - Example Real-world: Probability of 3 out of 10 mergers. Probability of there being 5 defective products in a batch of 20.
  - Formula for PMF: $\binom{n}{k} p^k (1-p)^{n-k}$
  - Formula for CDF is just the summation
  - It is more useful for small n's
  - Mean: np, variance: np(1-p)

EX :- Toss a coin n times, probability of k heads or k tails in ntrials

- **Poisson**
  - Discrete distribution that signifies the probability of 'x' occurrences of a certain event over a certain period of time or space.
  - Examples: Number of defaults per month, Number of banks per square kilometre.
  - PMF (not PDF) $\frac{\lambda^k}{k!} e^{-\lambda}$
  - Mean and variance are λ (lambda >0)

PMF is probability mask function

- **Geometric**
  - Number of attempts before an event
  - The interarrival distribution counterpart of a binomial. The coin toss case (uniform, binomial, geometric)
  - PMF $(1-p)^{k-1} p$
  - CDF $1-(1-p)^k$
  - Mean is $\frac{1}{p}$, and variance $\frac{1-p}{p^2}$

How many times I need to toss to get my first head/tail

- **Exponential**
  - The interarrival times of the Poisson distribution
  - The continuous version of the geometric distribution
  - Memoryless
  - PDF: $\lambda e^{-\lambda x}$, where lambda>0
  - CDF: $1\text{-}e^{-\lambda x}$
  - Mean: $\frac{1}{\lambda}$
  - Variance: $\frac{1}{\lambda^2}$

Example :-How long to wait for next person to arrive at toll

	Interarival Distribution	Count per unit interarrival distribution
Discrete Interarrival	Geometric	Binomial
Continuous interarrival	Exponential	Poisson

	Continuous Distribution
	Discrete Distribution

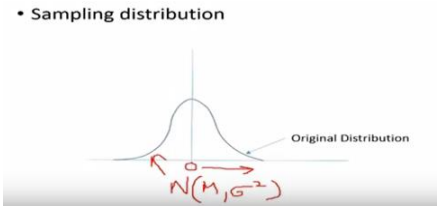
- Going from PDF to CDF (continuous)

$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

- Going from CDF to PDF (continuous)

$$f(x) = \frac{d}{dx}F(x)$$

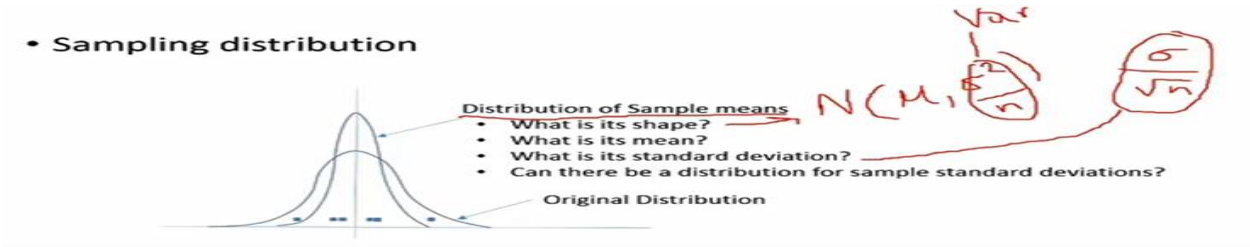
- Normal
  - Bell shaped curve
  - PDF: $f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
  - Mean, variance, CDF
  - Height, weight, etc.
  - Many things after removal of outliers
  - Binomial Approximation
  - Central Limit Theorem (CLT)
  - Sampling distributions



Central limit theorem

From original distribution take out sample and calculate Mean. For example I take random sample of each from set . The mean of all samples in aggreagation will be equal to overall mean of the distribution.

As we keep on increasing the sample the mean will tend to come close to Mean of the distribution. That is called Central Limit Distribution



Introduction to Inferential Statistics

- The overarching principle:
  - Have a null and alternate hypothesis
  - Do some basic calculations/arithmetic on the data to create a single number called the "test statistic"
  - If we assume the null hypothesis to be true (and make some assumptions about the distributions of various variables), then the 'test statistic' should be no different than a single random draw from a specific probability distribution.
  - Test the probability that the "test statistic" you calculated belongs to this theoretical distribution. This is the p-value!
  - Ergo: Its D|H not H|D

A **t-test** is used for **testing** the mean of one population against a standard or comparing the means of two populations if **you do not know the populations' standard deviation** and

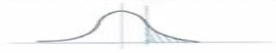
when you have a limited sample (n < 30). If **you know the populations' standard deviation**, you may use a **z-test**

Single sample z-test

- Using the rubric for this example:
  - Have a null and alternate hypothesis; $H_0: \mu_0 \leq 4.8$ and $H_{alt}: \mu_0 > 4.8$
  - Do some basic calculations/arithmetic on the data to create a single number called the "test statistic"; $z_{stat} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
  - If we assume the null hypothesis to be true (and make some assumptions about the distributions of various variables), then the 'test statistic' should be no different than a single random draw from a specific probability distribution. This is the Z-distribution or $N(0,1^2)$
  - Test the probability that the "test statistic" you calculated belongs to this theoretical distribution. This is the p-value!; Use Z-tables, Excel, Matlab or R
  - Low enough p-value is grounds for rejecting the null hypothesis

Single sample z-test

- The p-value is the probability of seeing a test statistic as extreme as the calculated value if the null hypothesis is true.
- If z_{stat} was computed to be 1.2 then
- P-value, Based on the standard null hypothesis:
- $H_0: \mu \leq 4.8$



- If null hypothesis was:
- $H_0: \mu \geq 4.8$

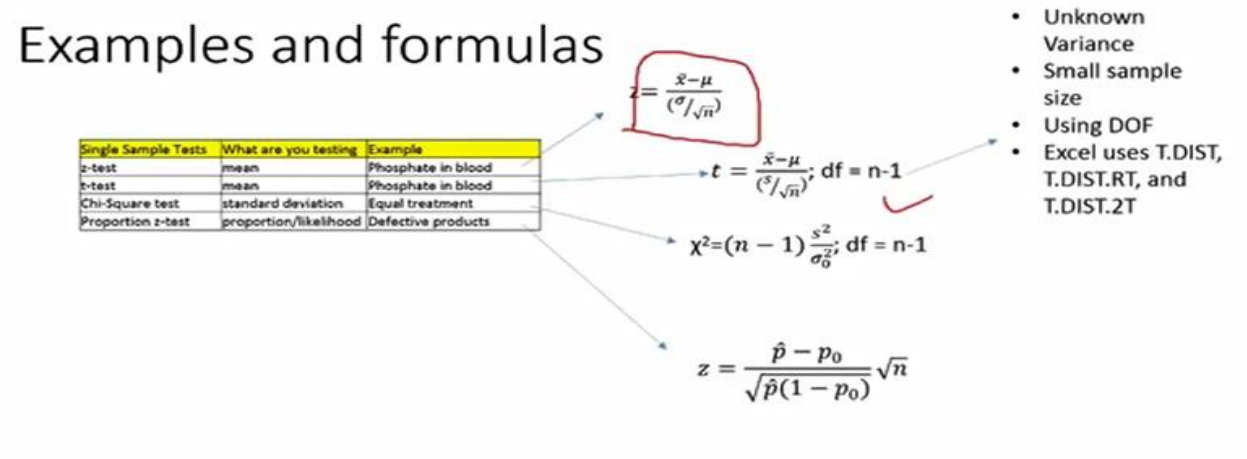
- $H_0: \mu = 4.8$ (two tailed)

- $H_0: \mu \leq 4.8$

- If null hypothesis was:
- $H_0: \mu \geq 4.8$

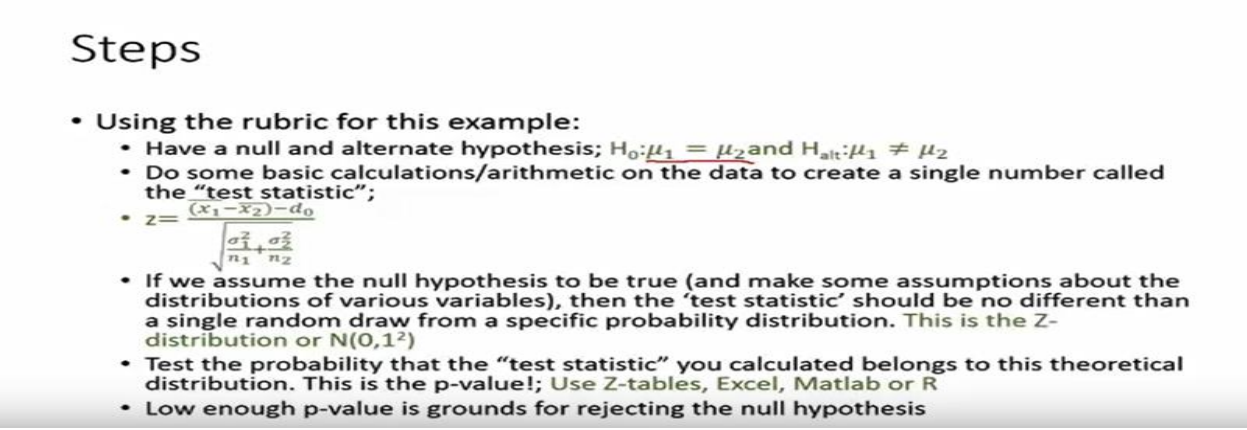
- $H_0: \mu = 4.8$ (two tailed)

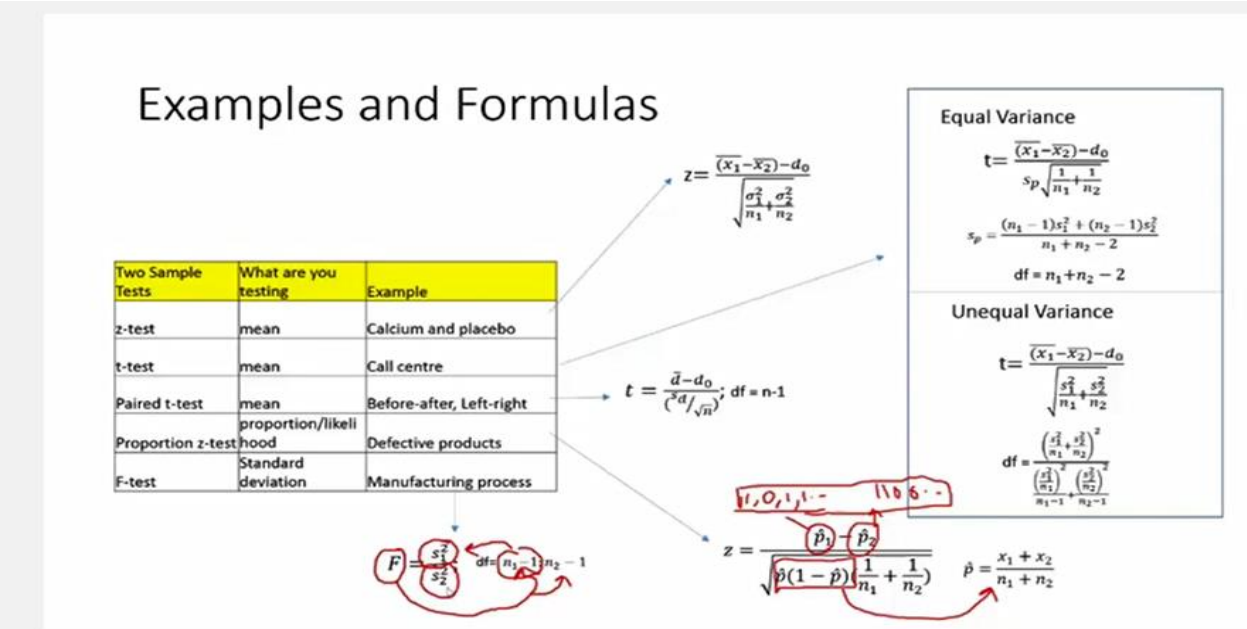
- Excel for right tail: =1-NORM.S.DIST(1.2,TRUE)
- Excel for left tail = NORM.S.DIST(1.2,TRUE)
- Excel for two-tail = 2*(1-NORM.S.DIST(1.2,TRUE))



T distribution :-Mean is 0,SD =1

Two Sample Tests





Acceptance Matrix

		Decision	
		Reject The Null Hypothesis	Fail to Reject the Null hypothesis
Actual	Null hypothesis is true	Type 1 Error (or Producer Risk, False Positive, alpha-risk)	Correct Decision (1-alpha)
	Alternate Hypothesis is true	Correct Decision (Power = 1-Beta)	Type 2 Error (Consumer risk, False Negative, Beta risk)

Pvalue is ideally type 1 error (alpha)

Type 2 error is ideally dependent on Alpha

CONFIDENCE INTERVAL:-

- Different ways of conceptualizing:
  - If we were to repeatedly take identical samples (same size) and build similar CI bounds for each sample then 95% of such CI bounds will cover the true mean.
  - We are 95% confident/certain that the true mean is within our confidence Interval.

Identify range when we don't have any hypothesis in place.

Examples and formulas

Single Sample Tests	What are you testing	Example
z-test	mean	Phosphate in blood
t-test	mean	Phosphate in blood
Chi-Square test	standard deviation	Equal treatment
Proportion z-test	proportion/likelihood	Defective products

$$z = \frac{\bar{x}-\mu}{(\sigma/\sqrt{n})} \rightarrow \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$t = \frac{\bar{x}-\mu}{(s/\sqrt{n})};\ df = n-1 \rightarrow \bar{x} \pm t_{\alpha/2,n-1}\frac{\sigma}{\sqrt{n}}$$

$$\chi^2 = (n-1)\frac{s^2}{\sigma_0^2};\ df = n-1 \rightarrow \left(\frac{S^2(N-1)}{\chi^2_{1-\alpha/2}}, \frac{S^2(N-1)}{\chi^2_{\alpha/2}}\right)$$

$$z = \frac{\hat{p}-p_0}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{n} \rightarrow \hat{p} \pm z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

ANOVA

- **Analysis of variance (ANOVA) is used as a test of means for two or more populations. The null hypothesis, typically, is that all means are equal.**

Statistics Associated with One-Way Analysis of Variance

- **eta^2 (η^2).** The strength of the effects of X (independent variable or factor) on Y (dependent variable) is measured by **eta^2 (η^2).** The value of η^2 varies between 0 and 1.
- ***F* statistic.** The null hypothesis that the category means are equal in the population is tested by an ***F* statistic** based on the ratio of mean square related to X and mean square related to error.
- **Mean square.** This is the sum of squares divided by the appropriate degrees of freedom.

Fstat = Mean Sqaure related to X/Mean square related to error

Y

$SS_y = SS_{between} + SS_{within}$

ANOVA OUTPUT

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-Stat
Between Treatments	$n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{\bar{y}}_{..})^2$ or SSB	a-1	MSB = SSB/DoF	F = MSB/MSE
Error within treatments	$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{i,j} - \bar{y}_{i.})^2$ or SSE	N-a	MSE = SSE/DoF	
Total	$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{i,j} - \bar{\bar{y}}_{..})^2$ or SST	N-1	MST = SST/DoF	

Compare F calculated against the F-distribution with a-1,N-a degrees of freedom and get a p-value

Its basically standard deviation within and between groups

Why F for difference in means??

- The F is the ratio of two variances (where the samples come from a normal distribution and the null hypothesis is that the variances are equal)
- MSB is a way of calculating total variance
- MSE is a way of calculating total variance
- MSB, MSE and MST will be equal if the null hypothesis is true
- However is the null hypothesis is not, then MSB>MST>MSE

What do you do after rejecting the null hypothesis

- Need to figure out which pairs of treatments are different. One popular way is the Tukey test
- Method 1:
  - Decide on an alpha value
  - Calculate the critical tukey distance with this formula:

$$T_\alpha = q_{\alpha,i,N-i}\sqrt{\frac{MSE}{n}}$$

,where i is the number of treatments, n is the number of replicates per treatment, and N is the total number of data points

  - Do a complete enumeration of all pairs of differences in means. i.e.: All possible ABS($yi,.- yj,.$)
  - See which of these are above the critical distance

Tukey Distance calculates the pairs which are different.

Chi-Square TOI

- When using categorical variables
- Use this to test:
  - Does the input categorical variable effect the output categorical variable (works 2 or more states of the input or output variable)
  - Independence between two variables
  - Construct a contingency table:

Exercise \ Smoking habit	Heavy	Regular	Occasional	Never
Frequent	7	9	12	87
Some	3	7	4	84
None	1	1	3	18

- Then if the null hypothesis is true then the test statistic is:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}$$

With (r-1)*(c-1) degrees of freedom (or rc-c-r+1)

Cross sectional :- not a function of time

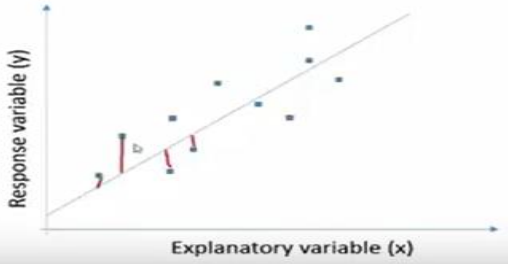
Time-series : previous time data impact current data. So the data attributes are function of time.

Linear Regression

- Minimize sum squared error
- With sufficient data simple enough
- With many dimensions, challenge is to avoid over fitting
  - Regularization
- Higher order functions?
  - Basis transformations
  - Ex: $(x_1, x_2) \rightarrow (x_1^2, x_2^2, x_1x_2, x_1, x_2)$

Ordinary Least Squares (OLS)

- Context:
  - Supervised Learning
- Derivation of OLS
  - Fit a line of the form y= mx+c or y = b_0+b_1x
  - Concept of actual y (y_i) and estimated y $(\hat{y}_i)$
  - Minimize the Squared deviation between actual and estimate. $\sum(y_i - \hat{y}_i)^2$



$$y_i = b_0 + b_1 x_i + e_i$$

$$e_i = y_i - b_0 - b_1 x_i$$

$$SSE = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Minimize the Squared deviation between actual and estimate. $(y_i - \hat{y}_i)^2$ and $\hat{y}_i = b_0 + b_1 x_i$

- Metrics to evaluate a Regression Model
  - We have so far discussed the p-value from the F-test of an ANOVA
  - What about R^2 and Adj R^2
  - For R^2 (It is nothing but SSM/SST)
    - From the ANOVA output SST, SSM/Regression, SSE/Residual

10	ANOVA					
11		df	SS	MS	F	Significance F
12	Regression	4	3555.051376	888.7628	100.6527	5.12691E-31
13	Residual	83	732.8892778	8.829991		
14	Total	87	4287.940654			
15						

    - SST = $\sum(y_i - \bar{y})^2$, SSM = $\sum(\hat{y}_i - \bar{y})^2$, SSE = $\sum(y_i - \hat{y}_i)^2$
  - Adj R^2 is nothing but: $1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$

Regularization Techniques

- Going beyond variable selection, what about variable shrinkage
  - Multicollinearity and the potential for many forms of a regression equation
  - Y = 4A-2B or Y = 10A-8B
- Ridge Regression
- $\hat{\beta}^{ridge} = \min \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$

Subject to $\sum_{j=1}^{p} \beta_j^2 \leq s$

- Lasso Regression
- $\hat{\beta}^{ridge} = \min \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$

Subject to $\sum_{j=1}^{p} |\beta_j| \leq s$

Regularization Techniques

- Going beyond variable selection, what about variable shrinkage
  - Multicollinearity and the potential for many forms of a regression equation
  - Y = 4A-2B or Y = 10A-8B
- Ridge Regression
- $\hat{\beta}^{ridge} = \min \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$

Subject to $\sum_{j=1}^{p} \beta_j^2 \leq s$

- Lasso Regression
- $\hat{\beta}^{ridge} = \min \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$

Subject to $\sum_{j=1}^{p} |\beta_j| \leq s$

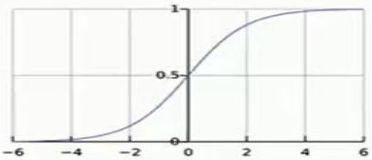
Regression for Classification

- What are the problems?
  - Linear regression is not limited in range
  - Output cannot be interpreted as a probability
    - Can be negative!
  - Works in practice, but not that well

- *Logistic or Logit* function
  - Log-Odds
  - Let $p(x)$ denote the $p(y=1|x)$
  - Logit transformation is given by: $\log\left(\frac{p(x)}{1-p(x)}\right)$

- Formally a logistic regression model tries to fit:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + x \cdot \beta_1$$



- Solving for $p(x)$

$$p(x) = \frac{e^{\beta_0 + x\cdot\beta}}{1 + e^{\beta_0 + x\cdot\beta}} = \frac{1}{1 + e^{-(\beta_0 + x\cdot\beta)}}$$

- It is the probability of the training data D, given a parameter setting
  - It is a function of the parameter, since the training data D is fixed

$$L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{(1-y_i)}$$

X