Stat 13 Final Review

- A. Probability tables to use.
- B. variance algebra, correlation, covariance, regression
- C. Probability and Conditional probability

Stat 13 Final review

A. Probability Tables to use.

Before (midterm) After

Normal distribution

Chi-square distribution

t-distribution

How to standardize?

Mean

Variance=

(Standard deviation)^2

Review lecture 3, especially slide 4

Degrees of freedom (d.f)

1. For one sample, d.f.=n-1(lecture 12, slide 9, lecture 13, 14)

2.For frequency/count, d.f.= number of cells -1 - number of parameters estimated (lecture 15,16,20/21)

3. For linear regression, d.f= sample size - 2 (lecture 25)

Pearson's chi-square

- Sum of

  (Observed - expected)^2/expected

For test of independence, degree of freedom equals (#Columns -1)(#rows -1)

Stat13 Final Review
Part B

Before (Midterm) After

Variance algebra, confidence interval

- Independent:

Var(X-Y)=var(X)
+var(Y)

- Dependent :

Var (X-Y)=
Var(X)+Var(Y)
-2Cov(X,Y)

Lecture 6,7

Regression line:

Slope equals

r [SD(Y)/SD(X)]

Where r is the correlation coefficient

Lecture 23,24,25

Standard error of the mean

Correlation= cov(X,Y)/SD(X)SD(Y)

Consistency : if use n-1 in doing SD, then use n-1 for averaging product

Practice: Step by step for Covariance,variance, and correlation coefficients.

x	y	X-EX	Y-EY	product	$(X-EX)^2$	$(Y-EY)^2$
2	4	-5	-1.5	7.5	25	2.25
4	3	-3	-2.5	7.5	9	6.25
6	6	-1	0.5	-0.5	1	0.25
8	5	1	-0.5	-0.5	1	0.25
10	8	3	2.5	7.5	9	6.25
12	7	5	1.5	7.5	25	2.25

E X=7 E Y=5.5 SD(X) =3.4 SD(Y)=1.7 Cov =29/6 Corr=0.828

sqrt(35/3)=3.4

Use population version, so divided by n

=cov/sd(x)sd(y)

Algebra for Variance, covariance

- $Var(X+Y) = Var\ X + Var\ Y + 2\ cov\ (X,Y)$
- $Var(X) = Cov\ (X, X)$
- $Var\ (X+a) = Var\ (X)$
- $Cov\ (X+a, Y+ b) = Cov(X,Y)$
- $Cov\ (aX, bY) = ab\ Cov(X,Y)$
- $Var(aX) = a^2\ Var\ (X)$
- $Cov(\ X+Y, Z) = cov(X,Z) + cov\ (Y,Z)$
- $Cov\ (X+Y, V+W) = cov(X,V) + cov\ (X, W) + cov\ (Y, W) + cov(Y,W)$

TRICK : pretend all means are zero;

$(X+Y)(V+W) = XV+XW+YW+YW$

Lecture 7 Accuracy of sample mean $\overline{X}$

Var ($\overline{X}$)= Var (X) divided by sample size n

What is X bar ? Called sample mean.

Standard error of the mean =SD($\overline{X}$)

= SD (X) divided by squared root of n

As sample size increases, the sample mean become more and more accurate in estimating the population mean

- Sample size needed to meet accuracy requirement

Stat 13 Final Review
Part C

- Probability function : mean and standard deviation; lecture 19,20,21
- Conditional probability : tree , table, should know how to update probability (Bayes theorem); lecture 17, 18

Binomial and Poisson

- You Should remember binomial
- $P(X=x)= \binom{n}{x} p^x (1-p)^{(n-x)}$

- I will provide Poisson in the exam; you should know how to use it
- $P(X=x)= e^{-\lambda} \lambda^x / x!$, where e =2.71828

Office hours next week
Monday, Wednesday 3-4pm

My office :

Geology 4608

Lecture 3 Normal distribution, stem-leaf, histogram

- Idealized Population, Box of infinitely many tickets, each ticket has a value.
- Random variable and probability statement P(X<85)
- Notations , Greek letters: Mean (expected value) and standard deviation, E(X) = μ , SD(X)= σ, Var(X)= σ^2
- Examples
- Empirical distribution : Stem-leaf, histogram
- Three variants of histogram : frequency, relative frequency, density(called “standardized” in book)
- Same shape with different vertical scale
- Density= relative frequency / length of interval

- Given a box of tickets with values that come from a normal distribution with mean 75 and standard deviation 15, what is the probability that a randomly selected ticket will have a value less than 85?

- Let X be the number elected (a random variable).
- Pr(X<85).

How does the normal table work?

- Start from Z=0.0 , then Z=0.1
- Increasing pattern observed
- On the negative side of Z
- Use symmetry

How to standardize?

- Find the mean
- Find the standard deviation
- Z= (X-mean)/SD

- Reverse questions:
- How to recover X from Z?
- How to recover X from percentile?

- Suppose there are 20 percent students failing the exam
- What is the passing grade?
- Go from percentage to Z, using normal table
- Convert Z into X, using X=mean + Z times SD

Probability for an interval

- P (60<X<85)
- Draw the curve (locate mean, and endpoints of interval)
- =P(X<85)-P(X<60) where
- P(X<60)= P(Z<(60-75)/15)=P(Z<-1)=1-P(Z<1)=1-.841= about .16

Lecture 12 Brownian motion, chi-square distribution, d.f.

- Adjusted schedule ahead
- Chi-square distribution (lot of supplementary material, come to class!!!) 1 lecture
- Hypothesis testing (about the SD of measurement error)and P-value (why n-1?supplement) 1 lecture
- Chi-square test for Model validation (chapter 11)
- Probability calculation (chapter 4)
- Binomial distribution and Poisson (chapter 5, supplement, horse-kick death cavalier data, hitting lottery, SARS infection)
- Correlation, prediction, regression (supplement)
- t-distribution, F-distribution

$R^2= (X_1-A)^2+ (X_2-A)^2 + \ldots + (X_n-A)^2$; $A= (X_1 + ..+X_n)/n$ =average

Follows a chi-square distribution with n-1 degrees of freedom

If variance of normal each X is σ^2

- Then D^2/σ^2 follows a chi-square distribution with n degrees of freedom
- R^2/σ^2 follows a chi-square distribution with n-1 degrees of freedom ; this is also true even if the mean of the normal distribution (for each X) is not zero (why?)

Lecture 13 Chi-square and sample variance

- Finish the discussion of chi-square distribution from lecture 12
- Expected value of sum of squares equals n-1.
- Why dividing by n-1 in computing sample variance?
- It gives an unbiased estimate of true variance of measurement erro
- Testing hypothesis about true SD of measurement error
- Confidence interval about the true SD of measurement error.

Measurement error= reading from an instrument - true value

One biotech company specializing microarray gene expression profiling claims they can measure the expression level of a gene with an error of size .1 (that is, after testing their method numerous times, they found the standard deviation of their measurement errors is 0.1) The distribution of errors follow normal distribution with mean 0 (unbiased).

Cells from a tumor tissue of a patient are sent to this company for Microarray assay. To assure consistency, the company repeat the assay 4 times. The result of one gene, P53 (the most well-studied tumor suppressor gene), is 1.1, 1.4, 1.5, 1.2.

Is there enough evidence to reject the company's claim about the accuracy of measurement? Note that sample SD is sqrt(0.1/3), Bigger than 0.1.

This problem can be solved by using chi-squared distribution. We ask How likely it is to observe a sample SD this big and if the probability is Small, then we have good evidence that the claim may be false . (next lecture)

Lecture 14 chi-square test, P-value

- Measurement error (review from lecture 13)
- Null hypothesis; alternative hypothesis
- Evidence against null hypothesis
- Measuring the Strength of evidence by P-value
- Pre-setting significance level
- Conclusion
- Confidence interval

- Testing statistics is obtained by experience or statistical training; it depends on the formulation of the problem and how the data are related to the hypothesis.
- Find the strength of evidence by P-value :

from a future set of data, compute the probability that the summary testing statistics will be as large as or even greater than the one obtained from the current data. If P-value is very small , then either the null hypothesis is false or you are extremely unlucky. So statistician will argue that this is a strong evidence against null hypothesis.

If P-value is smaller than a pre-specified level (called significance level, 5% for example), then null hypothesis is rejected.

Back to the microarray example

- H_o : true SD σ=0.1 (denote 0.1 by σ_0)
- H_1 : true SD $\sigma > 0.1$ (because this is the main concern; you don't care if SD is small)
- Summary :
- Sample SD (s) = square root of (sum of squares/ (n-1)) = 0.18
- Where sum of squares = $(1.1-1.3)^2 + (1.2-1.3)^2 + (1.4-1.3)^2 + (1.5-1.3)^2 = 0.1$, n=4
- The ratio s/σ_0 =1.8 , is it too big ?
- The P-value consideration:
- Suppose a future data set (n=4) will be collected.
- Let s be the sample SD from this future dataset; it is random; so what is the probability that s/ will be
- As big as or bigger than 1.8 ? $P(s/\sigma_0 > 1.8)$

- P(s/ σ_0 >1.8)
- But to find the probability we need to use chi-square distribution :
- Recall that sum of squares/ true variance follow a chi-square distribution ;
- Therefore, equivalently, we compute
- P (future sum of squares/ σ_0^2 > sum of squares from the currently available data/ σ_0^2), (recall σ_0 is
- The value claimed under the null hypothesis) ;

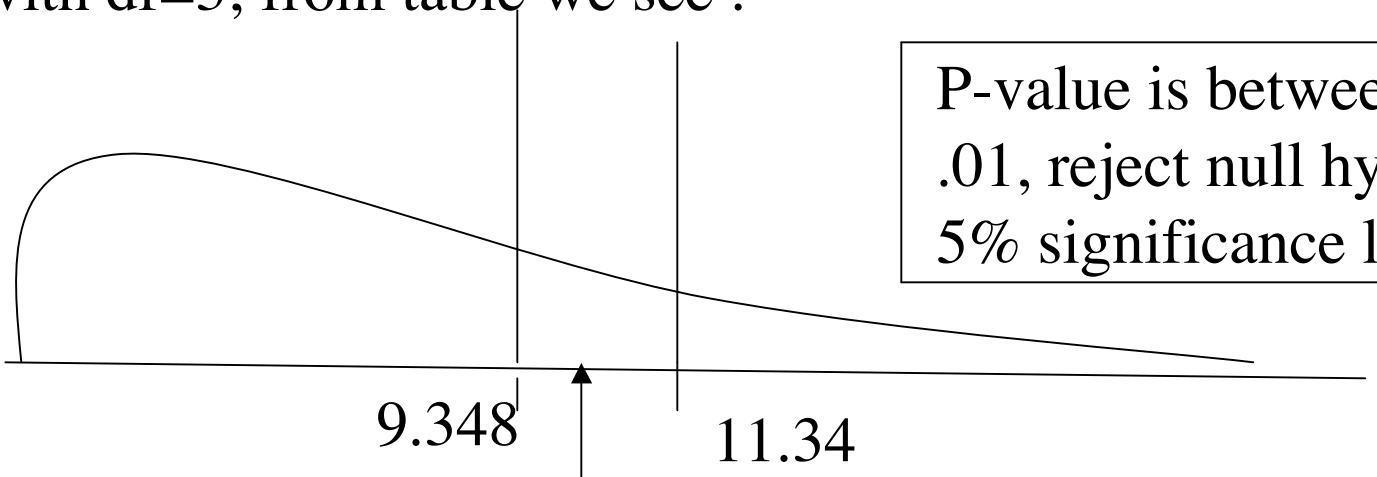
Once again, if data were generated again, then Sum of squares/ true variance is random and follows a chi-squared distribution

with n-1 degrees of freedom; where sum of squares= sum of squared distance between each data point and the sample mean

Note : Sum of squares= (n-1) sample variance = (n-1)(sample SD)^2

P-value = P(chi-square random variable> computed value from data)=P (chisquare random variable > 10.0)

For our case, n=4; so look at the chi-square distribution with df=3; from table we see :



P-value is between .025 and .01, reject null hypothesis at 5% significance level

The value computed from available data = .10/.01=10

(note sum of squares=.1, true variance =.1^2

Confidence interval

- A 95% confidence interval for true variance σ^2 is
- (Sum of squares/C_2, sum of squares/C_1)
- Where C_1 and C_2 are the cutting points from chi-square table with d.f=n-1 so that
- P(chisquare random variable > C_1)= .975
- P(chisquare random variable>C_2)=.025
- This interval is derived from
- P(C_1< sum of squares/ σ^2 <C_2)=.95

For our data, sum of squares= .1 ; from d.f=3 of table, C1=.216, C2=9.348; so the confidence interval of σ^2 is 0.1017 to .4629; how about confidence interval of σ ?

Lecture 15 Categorical data and chi-square tests

- Continuous variable : height, weight, gene expression level, lethal dosage of anticancer compound, etc --- ordinal
- Categorical variable : sex, profession, political party, blood type, eye color, phenotype, genotype
- Questions : do smoke cause lung cancer? Do smokers have a high lung cancer rate?
- Do the 4 nucleotides, A, T, G, C, occur equally likely?
-

Lecture 16 chi-square test (continued)

- Suppose 160 pairs of consecutive nucleotides are selected at random .
- Are data compatible with the independent occurrence assumption?

	A	T	G	C
A	15	10	13	7
T	10	13	7	10
G	10	10	10	10
C	5	12	10	8

Independence implies joint probability equals product of marginal probabilities

- Let P(first nucleotide = A)= P_{A1}
- P(first nucleotide = T)= P_{T1} and so on
- Let P (second nucleotide = A)= P_{A2}
- P(second nucleotide = T)= P_{T2} and so on
- P(AA)= P_{A1} P_{A2}
- P(AT) = P_{A1} P_{T2}
- We do not assume $P_{A1} = P_{A2}$ and so on

Expected value in ( ) ; df = (# of rows -1)(# of columns -1)

	A	T	G	C
A	15 (11.25)	10 (12.66)	13 (11.25)	7 (9.84)
T	10 (10)	13 (11.25)	7 (10)	10 (8.75)
G	10 (10)	10 (11.25)	10 (10)	10 (8.75)
C	5 (8.75)	12 (9.84)	10 (8.75)	8 (7.66)

Pearson's chi-square statistic= 166.8 > 27.88. P-value<.001

Simple or composite hypothesis

- Simple : parameters are completely specified
- Composite : parameters are not specified and have to be estimated from the data

- Loss of 1 degree of freedom per parameter estimated
- Number of parameters estimated = (# of rows -1)+
- (# of columns -1)
- So the df for chi-square test is #of cells -1 - (#of rows -1) - (# of columns -1) = (#of row -1)(#of col -1)

Test of independence in a contingency table

Are SARS death rates independent of countries ? Data from LA -times , as of Monday 5.pm. (Wednesday, from April 30, 2003)

	China	Hong Kong	Singapore	Canada	others
cases	3303	1557	199	344	243
death	148	138	23	21	11

Df = 1 times 4 = 4; but wait,

convert to death - alive table first

d.f. = 4

	China					total
death	148 (199.5)	138 (94)	23 (12)	21 (20.8)	11 (14.7)	341
alive	3155 (3103.5)	1419 (1463)	176 (187)	323 (323.2)	232 (228.3)	5305
total	3303	1557	199	344	243	5646

Pearson's Chi-square statistic =47.67 > 18.47; P-value<.001, reject null hypothesis, data incompatible with independence assumption

Lectures 20/21 Poisson distribution

- As a limit to binomial when n is large and p is small.
- A theorem by Simeon Denis Poisson(1781-1840). Parameter λ= np= expected value
- As n is large and p is small, the binomial probability can be approximated by the Poisson probability function
- $P(X=x)= e^{-\lambda} \lambda^{x} / x!$, where e =2.71828
- Ion channel modeling : n=number of channels in cells and p is probability of opening for each channel;

Binomial and Poisson approximation

x	n=100, p=.01	Poisson
0	.366032	.367879
1	.36973	.367879
2	.184865	.183940
3	.06099	.061313
4	.014942	.015328
5	.002898	.003066
6	.0000463	.000511
7		

Advantage: No need to know n and p; estimate the parameter λ from data

X= Number of deaths	frequencies
0	109
1	65
2	22
3	3
4	1
total	200

200 yearly reports of death by horse-kick from10 cavalry corps over a period of 20 years in 19th century by Prussian officials.

x	Data frequencies	Poisson probability	Expected frequencies
0	109	.5435	108.7
1	65	.3315	66.3
2	22	.101	20.2
3	3	.0205	4.1
4	1	.003	0.6
	200		

Pool the last two cells and conduct a chi-square test to see if Poisson model is compatible with data or not. Degree of freedom is 4-1-1 = 2. Pearson's statistic = .304; P-value is .859 (you can only tell it is between .95 and .2 from table in the book); accept null hypothesis, data compatible with model

Rutherfold and Geiger (1910)

- Polonium source placed a short distance from a small screen. For each of 2608 eighth-minute intervals, they recorded the number of alpha particles impinging on the screen

Other related application in

Medical Imaging : X-ray, PET scan (positron emission tomography), MRI

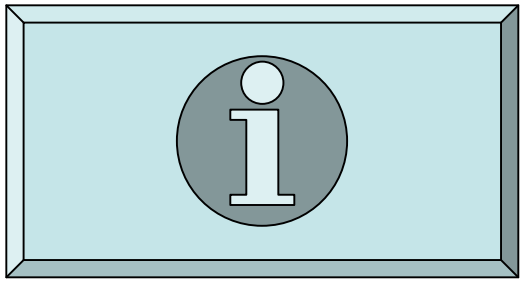
# of □ particles	Observed frequency	Expected freq.
0	57	
1	203	211
2	383	407
3	525	
4	532	508
5	408	394
6	273	254
7	139	140
8	45	68
9	27	29
10	10	11
11+	6	6

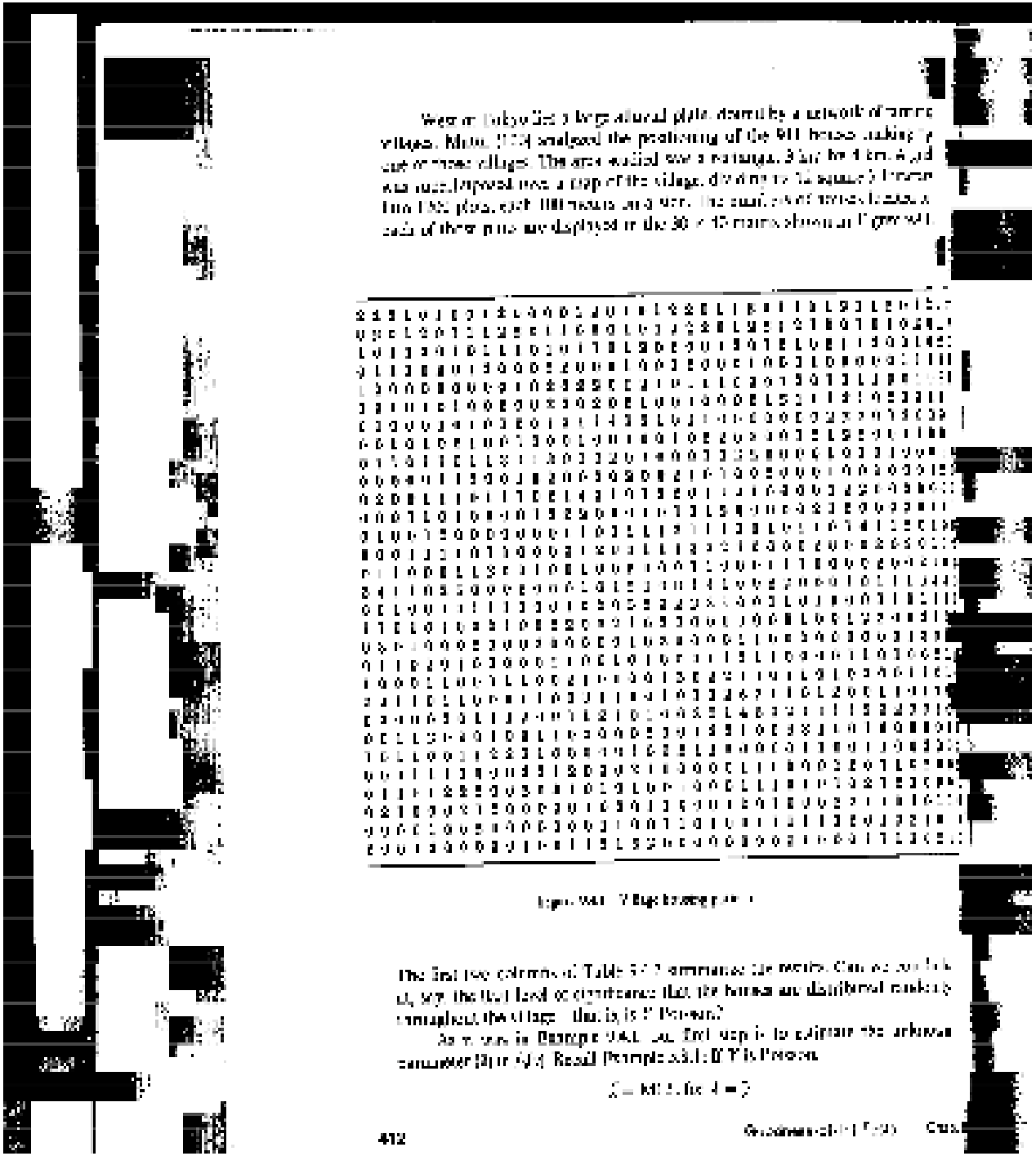
Pearson's chi-squared statistics = 12.955; d.f.=12-1-1=10

Poisson parameter = 3.87, P-value between .95 and .975. Accept null hypothesis : data are compatible with Poisson model

Poisson process for modeling number of event occurrences in a spatial or temporal domain

- Homogeneity : rate of occurrence is uniform
- Independent occurrence in non-overlapping areas
- Non-clumping





Stat13-lecture 25 regression (continued, SE, t and chi-square)

- Simple linear regression model:
- $Y = \beta_0 + \beta_1 X + \varepsilon$
- Assumption : ε is normal with mean 0 variance σ^2

  The fitted line is obtained by minimizing the sum of squared residuals; that is finding β_0 and β_1 so that

$(Y_1 - \beta_0 - \beta_1 X_1)^2 + \ldots. (Y_n - \beta_0 - \beta_1 X_n)^2$ is as small as possible

- This method is called least squares method

Least square line is the same as the regression line discussed before

- It follows that estimated slope β_1 can be computed by
- r [SD(Y)/SD(X)] = [cov(X,Y)/SD(X)SD(Y)][SD(Y)/SD(X)]
- =cov(X,Y)/VAR(X) (this is the same as equation for hat β_1 on page 518)
- The intercept β_0 is estimated by putting x=0 in the regression line; yielding equation on page 518
- Therefore, there is no need to memorize the equation for least square line; computationally it is advantageous to use cov(X,Y)/var(X) instead of r[SD(Y)/SD(X)]

Finding residuals and estimating the variance of ε

- Residuals = differences between Y and the regression line (the fitted line)
- An unbiased estimate of σ^2 is
- [sum of squared residuals]/ (n-2)
- Which divided by (n-2) ?
- Degree of freedom is n-2 because two parameters were estimated
- [sum of squared residuals]/σ^2 follows a chi-square.

Hypothesis testing for slope

- Slope estimate $\hat{\beta}_1$ is random
- It follows a normal distribution with mean

equal to the true β_1 and the variance

equal to σ^2 / [n var(X)]

Because σ^2 is unknown, we have to estimate from the data ; the SE (standard error) of the slope estimate is equal to the squared root of the above

t-distribution

- Suppose an estimate hat $\hat{\theta}$ is normal with
- variance c σ^2.
- Suppose σ^2 is estimated by s^2 which is related to a chi-squared distribution
- Then $(\hat{\theta} - \theta)/\sqrt{(c\ s^2)}$ follows a

t-distribution with the degrees of freedom equal to the chi-square degree freedom

An example

- Determining small quantities of calcium in presence of magnesium is a difficult problem of analytical chemists. One method involves use of alcohol as a solvent.
- The data below show the results when applying to 10 mixtures with known quantities of CaO. The second column gives
- Amount CaO recovered.
- Question of interest : test to see if intercept is 0 ; test to see if slope is 1.

X:CaO present	Y:CaO recovered	Fitted value	residual
4.0	3.7	3.751	-.051
8.0	7.8	7.73	.070
12.5	12.1	12.206	-.106
16.0	15.6	15.688	-.088
20.0	19.8	19.667	.133
25.0	24.5	24.641	-.141
31.0	31.1	30.609	.491
36.0	35.5	35.583	-.083
40.0	39.4	39.562	-.161
40.0	39.5	39.562	-.062

Standard error

Estimate

Least Squares Estimates:

- Constant -0.228090 (0.137840)
- Predictor 0.994757 (5.219485E-3)

- R Squared: 0.999780
- Sigma hat: 0.206722
- Number of cases: 10
- Degrees of freedom: 8

← Squared correlation

← Estimate of SD(ε)

.22809/ .1378 =1.6547

(1 - 0.994757)/ 5.219485E-3 = 1.004505233753904 4