Lecture Handout 6- Chapter 2 (continued) and statistics

Additional office hours this Friday after class until 11:30 AM and from 2:00 to 4:00 PM. Next week additional hours to be announced.

All lecture handouts, problem sets, past exams, and the exam formula sheet are on Compass. **Exam** will cover lectures, problem sets, past exams, items on Compass, book chapters 1 and 2, and book questions assigned.

Closed book exam; no notes allowed

Only the formula sheet devised by Dr. Roca will be allowed; this formula sheet will be included as a page on the exam (don't bring your own).

A copy of the formula sheet can be seen on Compass. I welcome suggestions for additions to the formula sheet.

Greek letter	Name	Roman letter	Statistical term
α	alpha	a	Type I error (see below)
β	beta	b	Type II error (see below)
δ	delta	D	difference
π	pi	p	proportion
μ	mu	M	mean
σ	sigma	s	standard deviation

Note: Greek letters are used to describe a population; Roman letters (i.e. English letters) are used to describe a sample

Statistics, review:

mean, or arithmetic mean: average of a group of values, and indicated by a bar above x

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Variance: measure of dispersion of values around a mean, sum of the SQUARES

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

TABLE 3–2

Calculation of the mean deviation

Column 1 number of coffee breaks X	Column 2 raw deviation $X - \bar{X}$	Column 3 absolute deviation $\lvert X - \bar{X} \rvert$	Column 4 squared deviation $(X - \bar{X})^2$
1	−8	8	64
3	−6	6	36
4	−5	5	25
7	[illegible]	[illegible]	4
9	0	0	0
9	0	0	0
11	2	2	4
12	3	3	9
16	7	7	49
18	9	9	81
$\Sigma X = 90$	$\Sigma(X - \bar{X}) = 0$	$\Sigma \lvert X - \bar{X} \rvert = 42$	$\Sigma(X - \bar{X})^2 = 272$

Why squared deviations? Just adding all the deviations from the mean always sums to zero; while absolute values are difficult to manipulate mathematically.

Standard deviation: another measure of dispersion, square root of the variance, so the value is on the same scale as the mean (i.e. not squared)

$$sd = (V_x)^{1/2}$$

Standard error: the standard deviation of a statistic, eg, of the mean. Note: 95% confidence interval, mean +/- 1.96(s.e.), will contain true mean 95% of the time

$$se = \left(\frac{V_x}{n}\right)^{1/2}$$

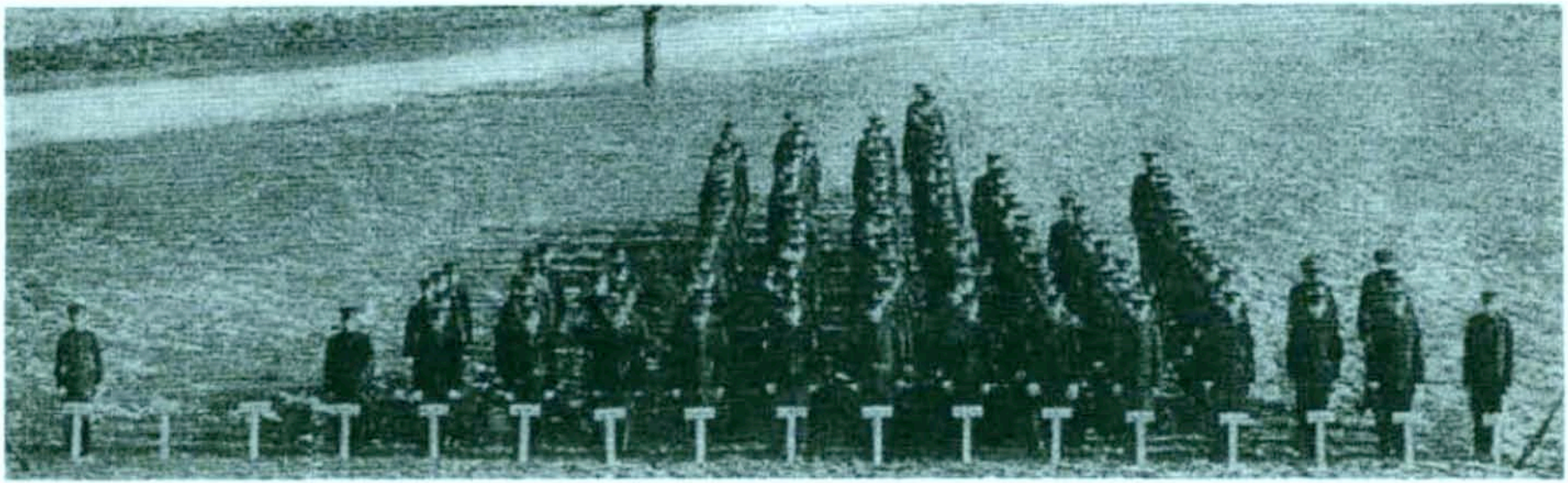
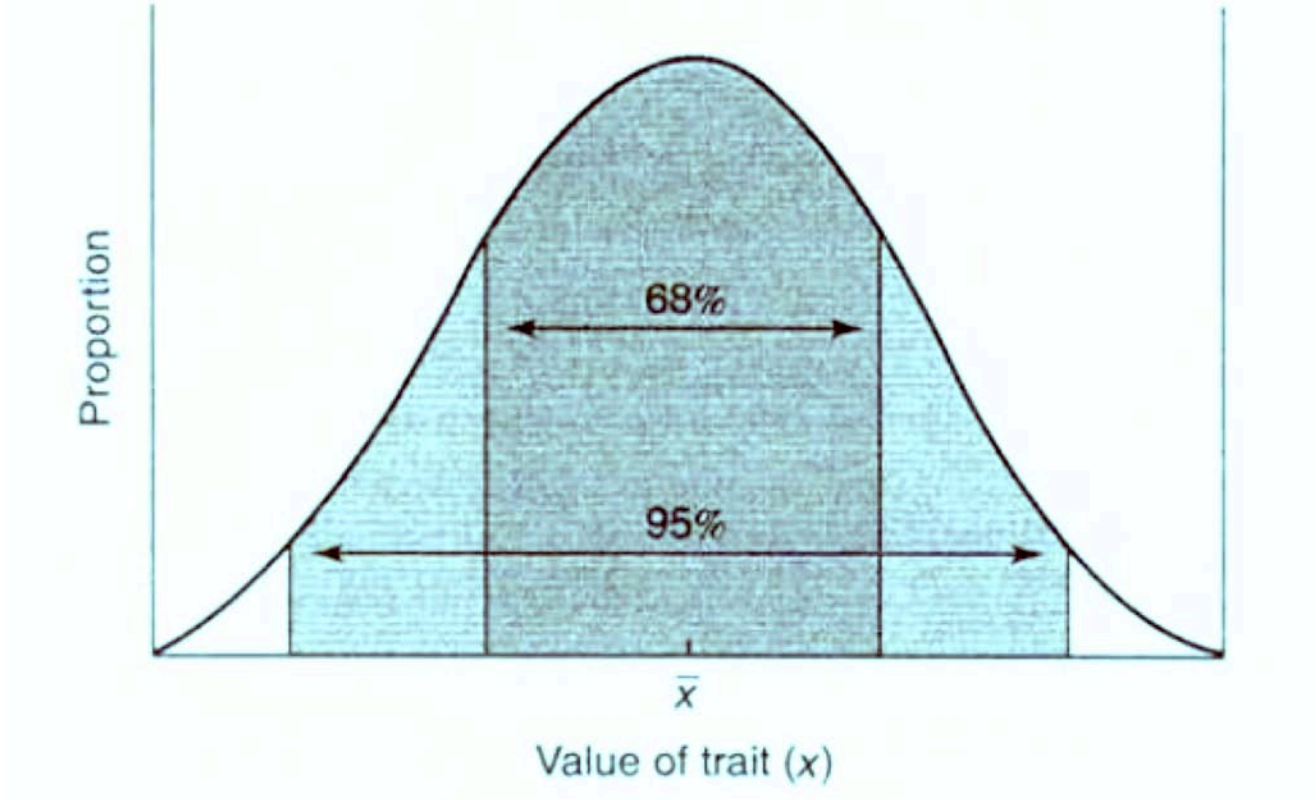


Figure 1.4. The distribution of heights, in inch categories, of male students at Connecticut Agricultural College (from Crow, 1997). (Courtesy of Albert Blakeslee, 1914.)

Normal distribution: often seen in samples from natural populations, in which values fall in a bell shaped curve in which 68% are within 1 sd of the mean, and 95% within 1.96 sd of the mean.



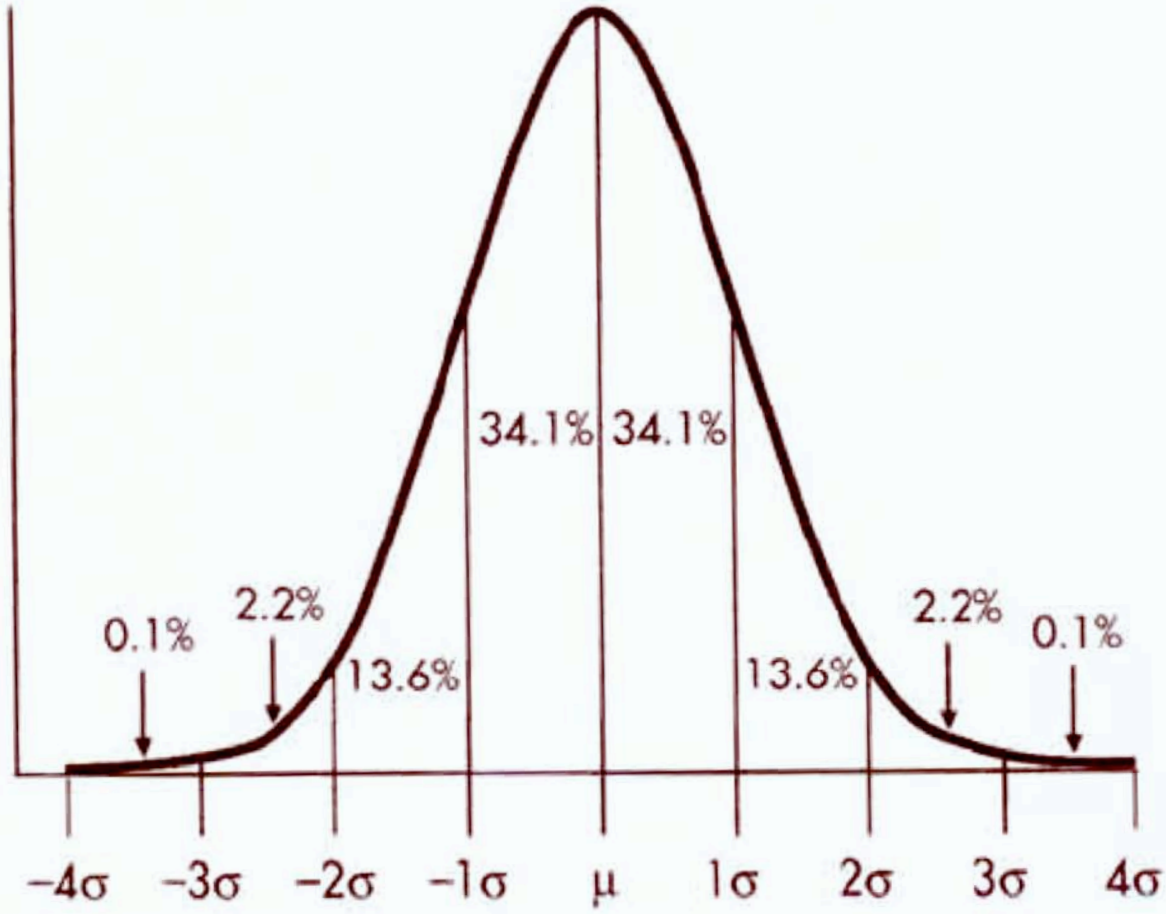


FIGURE 4–3
The normal curve.

TABLE 6–1

The relationship between α and β	Called	Truth: No difference	Truth: Difference
	Accept H_o	$(1 - \alpha)$	β
	Reject H_o	α	$(1 - \beta)$

For a null hypothesis: that a sample is not different from the population:

Alpha = Type I error = reject the null hypothesis even though it is true.

TABLE 6–1

The relationship between α and β	Called	Truth: No difference	Truth: Difference
	Accept H_o	$(1 - \alpha)$	β
	Reject H_o	α	$(1 - \beta)$

For the null hypothesis that a sample is not different from the population:

Beta = Type II error = accept the null hypothesis even though it is false.

TABLE 6–1

The relationship between α and β	Called	Truth: No difference	Truth: Difference
	Accept H_o	$(1 - \alpha)$	β
	Reject H_o	α	$(1 - \beta)$

For the null hypothesis that our sample matches the population, statistical **power** is the probability of concluding that there is a difference when in fact there is one.

Power = 1 - β

Please try in problem set: use of X^2 test to determine if allele frequencies are different between two populations. For two alleles:

$$\chi^2 = \frac{2N\,V(\hat{p})}{\bar{p}\bar{q}}$$

Where N is *combined* total sample size, p-bar and q-bar are *weighted* allele frequencies, and the weighted variance is:

$$V(\hat{p}) = \sum \frac{N_j}{N}\hat{p}_j^2 - \bar{p}^2$$

Estimates of allele frequency

-May not exactly reflect the allele frequency of a population if the sample size is small

-Larger sample sizes provide better estimates of population frequencies

-Sampling must not be biased

-Variance estimates reflect on the accuracy of allele frequency estimates

Estimates of allele frequency

-Generally, a maximum likelihood (ML) approach is used.

-Used to estimate the allele frequency p given the data. The maximum likelihood approach varies p until it finds the value that maximizes the probability of the data, given p.

ML approach, example: three genotypes at a codominant allele with sample size N, so N_{11}, N_{12}, N_{22}, with genotype frequencies p^2, $2p(1-p)$, and $(1-p)^2$

Multinomial probability to calculate the likelihood:

$$L = \frac{N!}{N_{11}!N_{12}!N_{22!}}(p^2)^{N_{11}}[2p(1-p)]^{N_{12}}[(1-p)^2]^{N_{22}}$$

$$= \frac{N!}{N_{11}!N_{12}!N_{22!}}2^{N_{12}}p^{2N_{11}+N_{12}}(1-p)^{N_{12}+2N_{22}} \qquad (2.10a)$$

The value of p that maximizes this probability also maximizes the probability of its logarithm. Therefore, we can calculate the logarithm

Five steps to calculate maximum likelihood estimation of gene frequency:

1. Determine likelihood (multinomial probability)
2. Transform with log
3. Take derivative
4. Set derivative to zero
5. Solve for allele frequency

Details are on page 80

SLIDES IN THIS HANDOUT
BEYOND THIS POINT ARE FOR
INFORMATION PURPOSES ONLY.

Why log(L) rather than L?

- Following power points were prepared by Byunggil Yoo, September 2007.

- Reminder: Maximum Likelihood is used to estimate the allele frequency p given the data. The maximum likelihood approach varies p until it finds the value that maximizes the probability of the data, given p.

Why log(L) rather than L?

- Likelihood (L) = $\frac{N!}{N_{11}!N_{12}!N_{22}!}2^{N_{12}}\,p^{2N_{11}+N_{12}}(1-p)^{N_{12}+2N_{22}}$

- Can we get Max(L) by dL/dp = 0 ?

  Yes, but there are up to $2N_{11}+2N_{12}+2N_{22}-1$

  values which satisfy dL/dp = 0 (L: polynomial)

- Why log(L) work?

  Because log is monotone increasing function,

  the value p which makes L maximum can also make log(L) maximum.

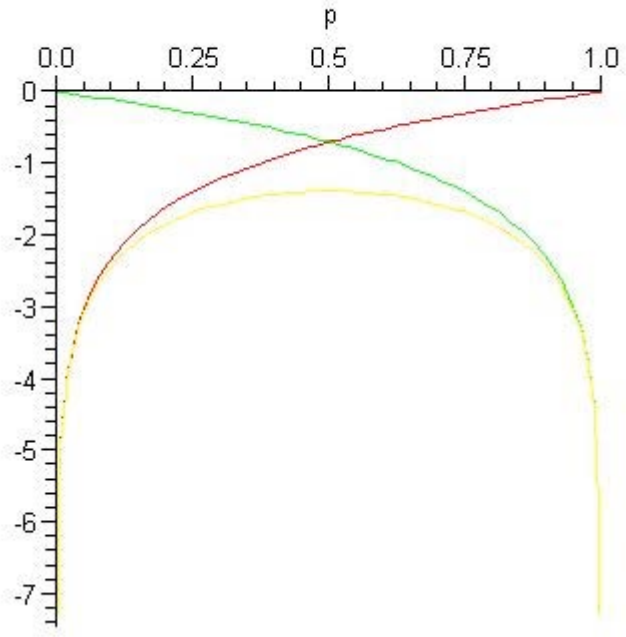
Advantages of log(L)?

- Can we get Max(log L) by d(log L)/dp = 0 ?
  Yes, in some special cases.

- If we can use d(log L)/dp = 0,
  Since Log (L) is much simpler than L, differentiation is easier, and number of values are considerably reduced.

Let’s plot

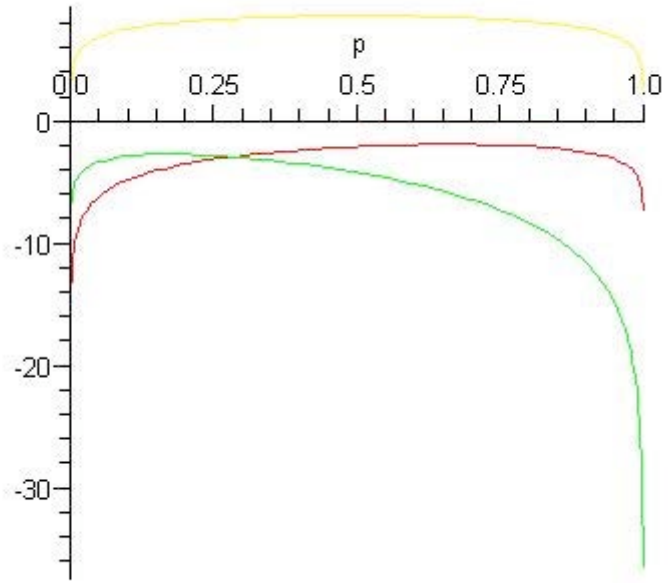
> plot({ln(p), ln(1-p), ln(p)+ln(1-p)}, p = 0 .. 1);



Sum of log(p) and log(1-p) result in a form which we can easily get Maximum by its derivative!

Influence of coefficient/constant?

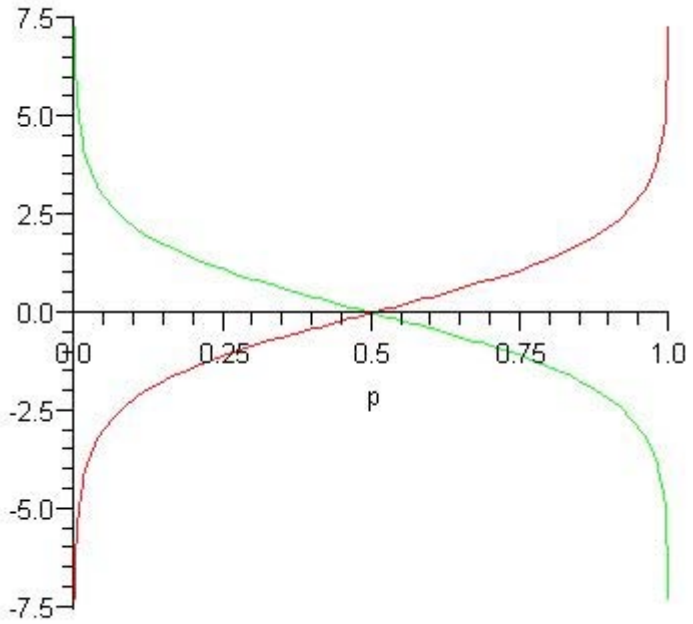
$$plot(\{\ln(p) + \ln(1-p) + 10,\ 2\ln(p) + \ln(1-p),\ \ln(p) + 5\ln(1-p)\},\ p = 0..1);$$



Those affects location of p which make log(L) maximum; p = 0 or 1 is always negative infinity, and there is ONLY one maximum!

Always work if Alog(p) + Blog(1-p)?

- No, Coefficient A, B should be positive!
- Example: $plot(\{\ln(p) - \ln(1-p), \ln(1-p) - \ln(p)\}, p = 0..1);$



In this case, A, B are always positive!