Course Outline

October 23, 2013
1:21 PM

Economic Statistics ECON 227
Instructor : K.MacKenzie Office : L435 Telephone : 514-398-4400 ext 00017
Textbook: (Paragraphe Book Store)
Statistics for Business and Economics (any edition), McClave, Benson and Sincich, Pearson/Prentice-Hall If you find second-hand earlier editions, they will be perfectly all right for the course.

- EDA (exploratory data analysis)
- Probability theory, including the expected value and standard deviation of random variables
- Geometric, binomial, Poisson, and hypergeometric discrete random variables
- Exponential and normal continuous random variables
- Sampling distributions
- Point and confidence-interval estimation. t and chi-square distributions
- One-population hypothesis tests
- Two-population tests
- F distribution
- ANOVA
- Chi-square tests of independence and goodness of fit
- Simple regression and correlation
- Multiple regression

Evaluation:

- assignments and quizzes 15
- midterm examination 25
- Final examination 60

Students should have a calculator capable of statistics computations with two-variable capacity.
There will be work done on the computer using MINITAB and EXCEL.
Work may be presented in English or French.
McGill University values academic integrity. Therefore all students must understand the meaning and consequences of cheating, plagiarism and other academic offences under the Code of Student Conduct and Disciplinary Procedures (see http://www.mcgill.ca/integrity for more information).

- Credit will be given for ONLY ONE of the following introductory statistics courses: AEMA 310, BIOL 373, ECON 227D1/D2, ECON 257D1/D2, GEOG 202, MATH 203, MGCR 271, MGCR 273, PSYC 204, SOCI 350.
- Credit will be given for ONLY ONE of the following intermediate statistics courses: AEMA 411, ECON 227D1/D2, ECON 257D1/D2, GEOG 351, MATH 204, PSYC 305, SOCI 461 with the exception that you may receive credit for both PSYC 305 and ECON 227D1/D2 or ECON 257D1/D2.
- If you have already received credit for MATH 324 or MATH 357, you will NOT receive credit for any of the following: AEMA 310, AEMA 411, BIOL 373, ECON 227D1/D2, ECON 257D1/D2, GEOG 202, GEOG 351, MATH 203, MATH 204, MGCR 271, MGCR 273, PSYC 204, PSYC 305, SOCI 350.

Lecture 13/09/09

October 23, 2013
1:20 PM

Exploratory Data Analysis – term coined by John Tukey

Mean symbol: x with bar over … (sigma) x / n

H-spread – same with quartiles

Lower hinge / upper hinge – median of lower half / median of upper half of data set

Lecture 13/09/11

October 23, 2013
1:19 PM

Exploratory Data Analysis – term coined by John Tukey

Mean symbol: x with bar over … (sigma) x / n

H-spread – same with interquartile range

Lower hinge / upper hinge – median of lower half / median of upper half of data set

$$H - spread = upper\ hinge - lower\ hinge$$

$$Upper\ inner\ fence = upper\ hinge + 1.5(upper\ hinge - lower\ hinge)$$

$$Lower\ inner\ fence = lower\ hinge - 1.5(upper\ hinge - lower\ hinge)$$

A number "x" in the data set is a Tukey-style is an outlier if x is bigger than the upper fence or smaller than the lower fence

Tukey's Five-Points

1. Lowest non-outlier = lower adjacent value
2. Lower hinge
3. Median
4. Upper hinge
5. Highest non-outlier = upper adjacent value

Example:

21	
23	
43	
45	
47	
47	Lower hinge
49	
50	
52	
56	
56	Median
57	
58	
59	
63	
64	Upper hinge
66	
71	
77	
82	
109	

H-spread:
Upper hinge – lower hinge
=64-47
=17

Upper inner fence:

Upper hinge + 1.5 (upper hinge – lower hinge)
=64 + 1.5(17)
=89.5

Lower inner fence:
Lower hinge – 1.5 (h-spread)
=47 – 1.5(17)
=21.5

1. Lowest non-outlier = lower adjacent value: 23
2. Lower hinge: 47
3. Median: 56
4. Upper hinge: 64
5. Highest non-outlier = upper adjacent value: 82

Box and whisker plot (insert diagram)

Percentile: a number x_p is the p^{th} percentile of a data set if

1) At least p percent of the data is less than or equal to it
2) At least 100 – p percent of the data is greater than or equal to it

First quartile – 25th percentile

25% of 22 is 5.5 -> Q1 = $n_5 + 0.5(n_6 – n_5)$ = 47

Third quartile – 75th percentile

75% of 22 is 16.5 -> Q3 = $n_{16} + 0.5(n_{17} – n_{16})$ = 65

Standardized data – z-score

If a data set has mean x-bar and standard deviation s_x then for a number x in the data set the z-score is:

$$\frac{x-\bar{x}}{s_x} = \frac{data - mean}{standard\ deviation} = z - score$$

Outliers have a z-score > 3

Lecture 13/09/16

October 23, 2013
1:18 PM

Probability

Probability questions assume perfect coins and perfect gender ratios (not necessarily what naturally occurs)

Monty Hall problem, 3 doors, 2 goats, 1 car.

Car	Pick	Open	Switch	
A	*A*	*B*	*N*	
A	*A*	*C*	*N*	
A	B	C	Y	
A	C	B	Y	
B	*B*	*A*	*N*	
B	*B*	*C*	*N*	
B	C	A	Y	
B	A	C	Y	
C	*C*	*A*	*N*	
C	*C*	*B*	*N*	
C	A	B	Y	
C	B	A	Y	

Experiment: any situation with an observable outcome.

Sample space: is a set of all possible outcomes of an experiment.

Event: a subset of the sample space.

Example:

Experiment – roll two dice

Sample space – all combinations of [1:6],[1:6]

$$\begin{pmatrix} 1,1 & \cdots & 6,1 \\ \vdots & \ddots & \vdots \\ 1,6 & \cdots & 6.6 \end{pmatrix}$$

The advantage of the 36-entry sample space is that all the outcomes are equally probable.

Let E be the event of rolling a 7. This event is a subset of the sample space.

If A implies B, A is a subset of B.

Lecture 13/09/23

October 23, 2013
1:17 PM
From last time

$$P(I_2) = P(I_2 \cap A_1) + P(I_2 \cap A_2) + P(I_2 \cap A_3) + P(I_2 \cap A_4)$$

Partition – the set of events (event 1, event 2, event 3) is called the partition of the sample space, provided two things are true: 1) e_i and e_j have no intersection $E_i \cap E_j = \varphi$ 2) if you take the union of all of them, equals the entire sample space $E_1 \cup E_2 \cup E_3 = the\ entire\ sample\ space$

Last class… Probability of colour blindness for the whole population:

P(n) = (0.02)(0.55)+(0.01)(0.45)

P(n) = 0.0155

Method II: Tree Diagrams

Method III: Joint Probability Table

	F	$\bar{F}$	
A	0.02 x 0.55 = 0.011	0.01 x 0.45 = 0.0045	0.056
$\bar{A}$	0.539	0.4455	0.944
	0.55	0.45	1.00

Example: Acme Inc. gets its USB keys from 3 suppliers: Able (40%), Baker (35%), Charles (25%) of USB keys. The defective rates are Able (0.2%), Baker (1.5%), Charles (4%).

a) What is the overall rate of defective rate of Acme's USB keys?

b) What proportions of the defective USB keys comes from the three suppliers?

Lecture 13/09/25

October 23, 2013
1:15 PM

Statistical Independence (Intuitive approach)

“Two events that have no impact on each other.”

It’s difficult to find two events that are strictly independent of each other.

Event A is said to be independent of Event B. The probability of A happening does not change for any change in Event B. This is true given 5 equivalent and basic criteria:

i) $P(A|B) = P(A|\bar{B})$

ii) $P(A|B) = P(A)$

iii) $P(A \cap B) = P(A)P(B)$ *

iv) $P(B|A) = P(B|\bar{A})$

v) $P(B)A = P(B)$

Works for everything, apparently.

Continuation:

i) implies iv)

i) $P(A|B) = P(A|\bar{B})$

Time out:

iv) $P(B|A) = P(B|\bar{A})$
Proof: Start with i)

$$P(A|B) = P(A|\bar{B})$$

$$\frac{P(A \cap B)}{P(B)} = \frac{P(A \cap \bar{B})}{P(\bar{B})}$$

$$\frac{P(A \cap B)}{P(B)} = \frac{P(A) - P(A \cap B)}{1 - P(B)}$$

$$P(A \cap B) - P(B)(P(A \cap B)) = P(A)P(B) - P(A \cap B)P(B)$$

$$P(A \cap B) = P(A)P(B)$$

iii) is true!

LS= $P(A|B) = \frac{P(B \cap A)}{P(A)} = \frac{P(A)P(B)}{P(A)} = P(B)$

RS= $P(B|A) = \frac{P(B \cap \bar{A})}{P(\bar{A})} = \frac{P(B) - P(B \cap A)}{1 - P(A)} = \frac{P(B) - P(A)P(B)}{1 - P(A)} = \frac{P(B)(1 - P(A))}{1 - P(A)} = P(B)$

In most cases iii) is easy to apply. Events A and B are statistically independent iff $P(A \cap B) = P(A)P(B)$. Important WARNING on when not to use iii): DO NOT calculate $P(A \cap B)$ as $P(A)P(B)$ unless it is certain A and B are independent.

The formula $P(A \cap B) = P(A)P(B)$ is used in two ways: 1) as a criterion for independence and 2) when independence is known, as a way to calculate the probability of $P(A \cap B)$.

Criterion for independence

Example:

	A	$\bar{A}$	
B	0.1	0.2	0.3
$\bar{B}$	0.4	0.3	0.7
	0.5	0.5	1.0

Solution:

$LS = P(A \cap B) = 0.1$

$LS = P(A)P(B) = (0.5)(0.3) = 0.15$

Since $(A \cap B) \neq P(A)P(B)$, A and B are statistically dependent.

Statistical independence is a purely mathematical criteria regardless of observed real world co-relation or causation.

Calculating $P(A \cap B)$

Example: A coin is tossed twice, let A be the event of heads on the first toss and B be the event of heads on the second toss. Let's supposed the coin is weighted, the probability of landing on heads is 60%. What is the probability of heads on both tosses?

$P(A \cap B) = P(A)P(B) = (0.6)(0.6) = 0.36$

What do we do for $P(A \cap B)$ if they are not independent?

Use the general multiplication rule!

$P(A \cap B) = P(A|B)P(B) = P(B \cap A) = P(B|A)P(A)$

$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$ ← Bayes' inversion rule

Practical Type Question: What is the probability in Eurelia of developing lung cancer if you are a smoker?

It's highly likely to find of lung cancer patients and the probability of the patient being a smoker.

In other words we are given: $P(S|C)$

$$P(C|S) = \frac{P(S|C)P(C)}{P(S)}$$

Random variables

A random variable associates a numerical value to each outcome of an experiment.

The value can be given in some natural way, or if necessary, in some artificial manner.

Random variables can be discrete or continuous (or a mixture of the two).

A discrete random variable means there will be gaps between successive possible values (ie. 1, 2, 3, but no 1.24, 3.141592654).

A continuous random variable can take all possible numerical values in some complete interval of values with no gaps.

Example: number of students present in a class is a discrete variable. Stepping on a scale and looking at a scale and looking at your ~~weight~~ mass. While it is true given a fine enough division, there will be jumps in mass, these are conceptually continuous variable.

Example: Let G = 1 if a voter is Female, 0 if a voter is male. This variable is discrete and artificially defined.

Lecture 13/09/30

October 23, 2013
1:14 PM

Random variables

A random variable associates a numerical value to each outcome of an experiment.

The value can be given in some natural way, or if necessary, in some artificial manner.

Random variables can be discrete or continuous (or a mixture of the two).

A discrete random variable means there will be gaps between successive possible values (ie. 1, 2, 3, but no 1.24, 3.141592654).

A continuous random variable can take all possible numerical values in some complete interval of values with no gaps.

Example: number of students present in a class is a discrete variable. Stepping on a scale and looking at a scale and looking at your ~~weight~~ mass. While it is true given a fine enough division, there will be jumps in mass, these are conceptually continuous variable.

Example: Let G = 1 if a voter is Female, 0 if a voter is male. This variable is discrete and artificially defined.

Discrete Random Variables

Concert Ticket Sales

Price	Probability
25	0.30
60	0.60
125	0.10

When the frequencies are probabilities, always use σ_x instead of s_x (see dentist problem in homework assignment)

Mean = $56.00

S.D. = $27.82

X	P(x)
25	0.30
60	0.60
125	0.10

This is called the pdf (probability distribution function) of the discrete random variable x
(see formulas for calculating S.D. for theory otherwise, always use calculator in this course)

$$\sigma_x = \sqrt{\sum (x_i - E(x))^2 P(x_i)}$$

For our current random variable x,

$$E(x) = 25(0.3) + 60(0.6) + 125(0.1)$$

$$E(x) = 56\,; as\ expected$$

Quiz is up to tree diagram material.

Histograms

Apparently not much explanation is needed

On a histogram without frequencies, assign arbitrary proportions and hope for the best.

Basic Counting Principle (singular)

If there are n_1 ways of doing the first thing, and for each of these ways, there are n_2 ways of doing the second thing, then there are $n_1 n_2$ ways of doing the two things. This principle extends to sequences of more than 2 actions.

Rationalizing Example:

I have 3 shirts and 2 neckties. How many different shirt-necktie ensembles can I wear? 6.

Solution, Brute-Force:

No clever thinking, just work out every possibility.

S1T1, S1,T2

S2T1, S2,T2

S3T1, S3,T2 = 6 combinations.

Solution, Clever:

Basic counting principle: There are three ways to choose the shirt and for each of these ways there are two ways to choose the tie:

n1n2, n1=2, n2=3, n1n2=(2)(3)=6

Combinations and Permutations

Some easy examples

1. How many 4-letter strings (sequences) are possible with a 26 letter alphabet? $26^4 = 456976$

2. How many 4-letter strings (sequences) are possible with a 26 letter alphabet with no repeated letters? $26 \times 25 \times 24 \times 23 = 358800$

3. I have 5 volumes of a 20-volume encyclopedia. In how many ways can I arrange all 5 books in a row? $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

$x!$ gives the number of ways of arranging x distinguishable items in a row

Eurelian Lottery

President KMacK visited Montreal 2 years ago. What he noticed was that people were buying tickets of 2 or 3 dollars for pieces of paper. When we collect a lot of it, we keep most of it and give the rest away at once in a large sum. So there are 49 numbers and you pick 6 out of the 49 numbers and if your match you win the large sum. (see PDF on mycourses)

Eurelian lottery rules: 3 different letters chosen. Different orders of the same letters are considered different tickets.

Brute Force: don't try this at home

26x25x24=15600

Actual way of doing this, take 15600 and divide it by 3!=6

Lecture 13/10/09

October 23, 2013
1:12 PM

Binomial Probability Formula

$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$

$E(x) = np$

$\sigma(x) = \sqrt{npq}$

Calculus proves this.

Example:

AJ Jones has a phone number that resembles the number of a pizza delivery restaurant. His estimate is that every time his phone rings there's a 15% probability that it's for the restaurant. In a random sample of 20 of Jones' calls monitored by the phone company a record will be kept by the company of how many of the calls are for pizza.

a) What is the probability that exactly three of the calls will be for pizza?
b) At most 3 are for pizza
c) At least 3 are for pizza

Solution:

a) $p(3) = \binom{20}{3} 0.15^3 (0.85)^{17} \cong 0.243$
b) *Make sure to include the possibility of 0*

$p(0) + p(1) + p(2) + p(3)$

$= \binom{20}{0} 0.15^0 (0.85)^{20} + \binom{20}{1} 0.15^1 (0.85)^{19} + \binom{20}{2} 0.15^2 (0.85)^{18} + \binom{20}{3} 0.15^3 (0.85)^{17} \cong 0.648$

c) $p(3) + p(4) + p(5) + p(6) + \cdots + p(20)$ or $1 - p(0) - p(1) - p(2)$

$= 0.595$

Hyper-geometric Introduction

20 women and 10 men are in a society. If 5 members are selected at random what is the probability that 3 of them will be women and 2 men.

$\frac{\binom{20}{3}\binom{10}{2}}{\binom{30}{5}} \cong 0.35998484 \cong 36\%$

2000 F 1000 F-compliments

$\frac{\binom{2000}{3}\binom{1000}{2}}{\binom{3000}{5}} \cong 0.3294 \cong 33\%$

Lecture 13/10/16

October 23, 2013
1:11 PM

What we have seen?

Binomial

Hyper-geometric

Grey area: two formulas give the same answer when the numbers are very large

Example: Forensic statistics.

Suppose a contract has 200 invoices, of which 4 have errors. How large a sample would have to be taken in order for the probability of finding at least one of the erroneous invoices in the sample exceeding 95%? The government allows 2% error in invoices.

Solution: This is a classic hyper-geometric situation. Let n be the sample size. P(x>=1) > 95%

$P(x \geq 1) > 95\%$

$P(1) + P(2) + P(3) + P(4) > 0.95$

$1 - P(0) > 0.95$

$$1 - \frac{\binom{4}{0}\binom{196}{n}}{\binom{200}{n}} > 0.95$$

$$0.05 > \frac{\binom{196}{n}}{\binom{200}{n}}$$

Guess and check n=105 P=0.049

b) If the binomial formula is used, are we in the grey area?

$P(0) < 0.05$

$$\binom{104}{n}(0.02)^0(0.88)^{104} = 0.1223$$

A Glimpse of the Future

Siméon Poisson came up with the Poisson distribution.

Motivational example: A certain 10 km stretch of highway is notorious for accidents. There is an average of 4.45 accidents per week. Accidents are equally likely to happen anywhere along the 10km of the highway. A week is selected at random. What is the probability of exactly 3 accidents in that week (independent)?

First estimate of the answer:

Think of the 10 km as a sequence of 10,000 meter long segments. For the time being, treat the probability of more than 4 (independent) accidents occurring within the same metre long segment as negligible. Then we say every segment either has an accident or doesn't. Having an accident is a success and the failure is not having an accident. Since accidents are equally likely everywhere, the P of success is equal across all n=10000 segments.

$mean = 4.45$

$np = 4.45$

$$p = \frac{4.45}{n} = \frac{4.45}{10000}$$

$$p(exactly\ 3\ accidents) = \binom{10000}{3}\left(\frac{4.45}{10000}\right)^3\left(1 - \frac{4.45}{10000}\right)^{9997} = 0.171529103$$

Prove assumption that P of more than one accident happening in the same segment is negligible, will try n= 1000000 instead.

$$p = \frac{4.45}{n} = \frac{4.45}{1000000}$$

$$p(exactly\ 3\ accidents) = \binom{1000000}{3}\left(\frac{4.45}{1000000}\right)^3\left(1-\frac{4.45}{1000000}\right)^{9997} = 0.171521487$$

Poisson's formula

$$\lim_{n\to\infty}\binom{n}{x}\left(\frac{\mu}{n}\right)^x\left(1-\frac{\mu}{n}\right)^{n-x}$$

l'Hôpital's rule gives:

$$\frac{\mu^x}{x!}e^{-\mu}$$

Lecture 13/10/21

October 21, 2013
3:53 PM

Poisson's formula

$$\lim_{n\to\infty}\binom{n}{x}\left(\frac{\mu}{n}\right)^x\left(1-\frac{\mu}{n}\right)^{n-x}$$

$$\mu = np = mean = 4.45$$

l'Hôpital's rule gives:

$$\frac{\mu^x}{x!}e^{-\mu}$$

Poisson set-up

1. A region of space of a length of time is involved
2. The number of occurrences of something is counted
3. An occurrence is equally likely to happen anywhere in the region of space or anytime during the length of time
4. $P(x) = \frac{\lambda^x}{x!}e^{-\lambda}$
   $\lambda = the\ mean\ number\ of\ occurences$
   $x = the\ exact\ number\ of\ occurences$

Example: At a beehive the number of arrivals of bees at the hive has approximately a poisson distribution with a mean of 2.8 per minute.

a) What is the probability of at least 3 arrivals in a randomly selected minute?

b) What is the probability of no arrivals in:

i) 30 seconds?

ii) 1.3 minutes?

c) A beekeeper starts observing the entrance to the hive at 9am. What is the probability of having to wait more than 45 seconds before the first bee arrives?

Solution: a)

$$P(x \geq 3) = P(3) + P(4) + P(5) + P(6) \ldots$$
$$P(x \geq 3) = 1 - P(0) - P(1) - P(2)$$
$$P(x \geq 3) = 1 - \frac{\lambda^0}{0!}e^{-2.8} - \frac{\lambda^1}{1!}e^{-2.8} - \frac{\lambda^2}{2!}e^{-2.8}$$
$$P(x \geq 3) = 0.538$$

b. i)

$$\lambda = \frac{1}{2}(2.8) = 1.4$$
$$P(0) = \frac{1.4^0}{0!}e^{-1.4} = 0.2466$$

b. ii)

$$\lambda = 1.5(2.8) = 4.2$$
$$P(0) = \frac{4.2^0}{0!}e^{-4.2} = 0.015$$

c)

$\lambda = 0.75(2.8) = 2.1$
$P(W > 45s)$
$= P(0\ arrivals\ in\ 45s) = \frac{2.1^0}{0!} e^{-2.1} = 0.1224$

d)

$P(W > t\ minutes)$
$= P(x = 0\ in\ t\ minutes)$
$\lambda = 2.8t$
$= \frac{2.8t^0}{0!} e^{-2.8} = e^{-2.8}$

Continuous formula for a random variable

$P(W > t) = e^{-\lambda t}$

W is the passage of time so it is at least conceptually continuous.

Counting large numbers of things

A swarm of bees is depicted in a drawing by KMacK.

Question: How many bees are here in the picture, supposing that the bees have landed randomly on the area depicted?

Method: Superimpose a 20x20 grid on the picture. Suppose there are 27 empty cells.

Let the number of bees be n.

For any particular cell the probability for every bee of ending up in that cell is $\frac{1}{400}$.

$P(0\ bees\ in\ a\ cell)$
$= \binom{N}{0} \left(\frac{1}{400}\right)^0 \left(1 - \frac{1}{400}\right)^N$
$= \left(\frac{399}{400}\right)^N = 0.9975^N \cong \frac{27}{400}$
$0.9975^N \cong 0.0675$
$N ln 0.9975 \cong ln 0.0675$
$N = \frac{ln 0.0675}{ln 0.9975} = 1077$

Method II: Poisson

$np = mean$
$n\left(\frac{1}{400}\right) = mean$
$P(0) \cong 0.0675$
$\frac{\lambda^0}{0!} e^{-\lambda} \cong 0.0675$
$e^{-\frac{n}{400}} \cong ln 0.0675$
$n \cong 1078$

Lecture 13/10/23

October 23, 2013
3:20 PM

So far:

Binomial-Hypergeometric

Binomial-Poisson

Grey areas are defined for Hypergeometric and binomial theorems as:

Works best when N_1 and N_2 are in the hundreds or bigger. Also works better if x is small ($x < 20$)

Grey areas for Binomial-Poisson

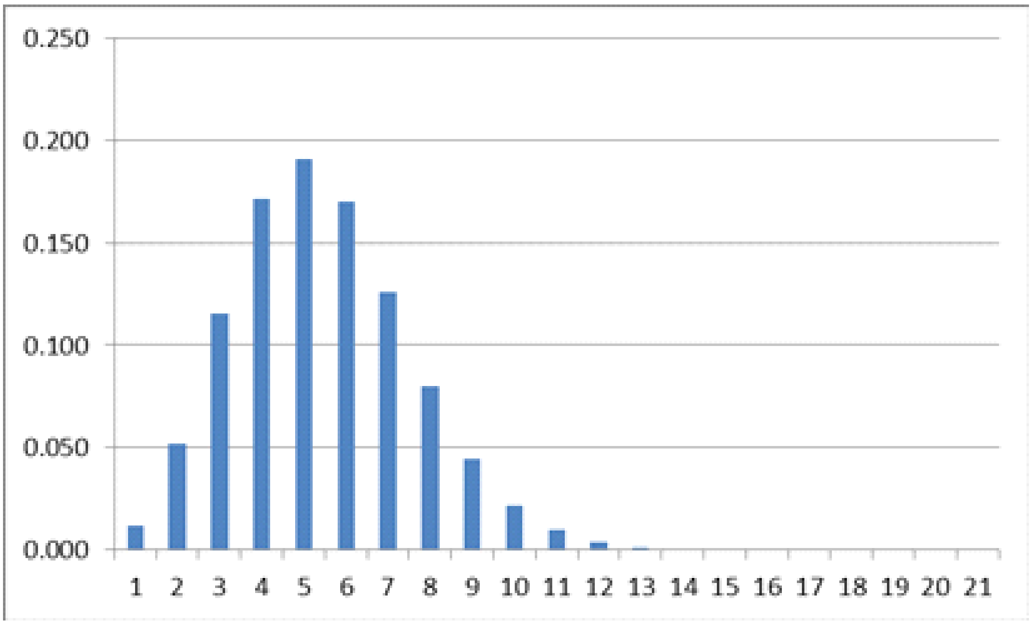
$p = \frac{\lambda}{n}$; $\lambda = np$

Works best if p is close to 0, some say $np < 7$ some say $np < 5$

More on Poisson

Write out a few entries of the PDF for a Poisson random variable with $\lambda = 4.45$

x_i	$P(x_i)$		Cumulative
0	0.012	$P(0) = \frac{4.45^0}{0!} e^{-4.45}$	0.012
1	=0.012 x 4.45 =0.052		0.064
2	=0.052 x 4.45 / 2 =0.116		0.179
3	=0.116 x 4.45 / 3 =0.172		0.351
4	=0.172 x 4.45 / 4 =0.191		**0.542**
5	=0.191 x 4.45 / 5 =0.170		0.711
6	=0.170 x 4.45 / 6 =0.126		0.837



The median in a PDF is the smallest value of the random variable for which the cumulative probability is 0.5 or greater. This Poisson distribution is slightly positively skewed since the mean (λ=4.45) is slightly bigger than the median (=4).

Example of median chasing:

We take our unfair coin that has p=0.6 of success and pbar=0.4. We toss the coin 12 times, and compare the mean and median.

$p = 0.6; n = 12;\ \lambda = 7.2$

$P(x) = \binom{12}{x}(0.6)^x(0.4)^{12-x}$

PDF:

x_i	$P(x_i)$		Cumulative
0	0.000017	$P(0) = \binom{12}{0}(0.6)^0(0.4)^{12-0}$	0.000017
1	0.000302		0.000319
2	0.00249		0.002809
3	0.01246		0.015269
4	0.0420		0.057269
5	0.1009		0.158169
6	0.1766		0.334769
7	0.2270		**0.561769**

Mean is greater than median, therefore there is a positive skew

Negative binomial theorem – *An amusement*

Suppose the confidence trickster uses our fake coin. Find the probability that the 4th head will occur on the 7th toss.

There must have been exactly 3 heads in 6 tosses then a head on the 7th toss.

$$P(3\ heads\ in\ 6\ tosses, followed\ by\ heads) = \left(\binom{6}{3}(0.6)^3(0.4)^3\right)(0.6)$$

Lecture 13/10/28

October 28, 2013
2:39 PM

Warm-up problem

Every time AJ Jones orders a coffee there's a 63% probability that the waiter will forget to bring a spoon. This is because Jones likes to steal the spoon and the restaurant has learned that he steals them. If we keep track of Jones' coffee orders what is the probability that he will need to order 10 cups to collect 3 spoons?

Method: before reaching the 10th cup, Jones needs to have exactly 2 spoons for the first 9 cups and then a spoon on the 10th cup.

$$\binom{9}{2}(0.37)^2(0.63)^7(0.37) = 0.0718$$

Official formula for negative binomial probabilities

1. Binomial set up applies - two things can happen success/failure, probability of success is the same every time.
2. It is desired to get the probability that the kth success will occur on the xth trial.

$$P(x) = \binom{x-1}{k-1}(p)^{(k-1)}(1-p)^{(x-k)}(p)$$
$$P(x) = \binom{x-1}{k-1}(p)^{(k)}(1-p)^{(x-k)}$$

Geometric distribution
A special case of the negative binomial theorem

Question: what is the probability of having to try x times in order to get the first success.

What is the probability that Jones will have to order 4 cups in order to get the first spoon?

$$P(x=4) = qqqp = (0.63)^3(0.37)$$
$$P(x) = (q)^{(x-1)}p$$

$$E(x) = \frac{1}{p}$$

That's it for the conventional discrete random variables- binomial, hypergeometric, poisson, negative binomial, geometric! See you next time!

Continuous Random Variables
It's KMacK's birthday!

Recall that for a discrete random variable (x) we can give PDF which is the set of possible values together with their probabilities.

Can we do the same for a (conceptually) continuous random variable?

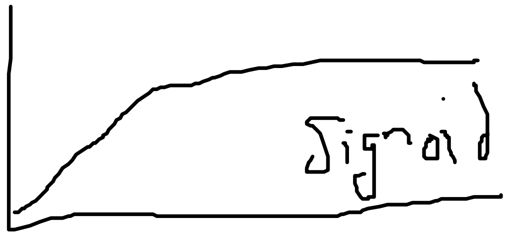
Let H be the height in cm of a randomly-selected adult Canadian male. What is the probability that the height = 175 *exactly.*

Is it 0 or 1/(infinity)? KMacK says 0.

What would the kind of PDF used for discrete random variable look like for a continuous random variable?

y_i	$P(y_i)$

y_i	$P(y_i)$
2.0312	0
-1.12381	0
887347.3	0



This table is useless.

Since that kind of PDF is not useful we use a CDF instead - a cumulative distribution function

Graph of the CDF has an asymptote at 1 and another at 0.

Lecture 13/10/30

October 30, 2013
2:39 PM

Continuous Random Variables

$P(x = k\ exactly) = 0;\ (aka \frac{1}{\infty})$

CDF: $F(t) = P(x \le t)$ A cumulative distribution function

- Not all CDFs are sigmoid curves
- The calculus derivative of the CDF is called the density function f(t)
  $f(t) = F'(t)$

Our first CDF

Let us suppose that the arrivals of taxis by the Roddick Gates have a Poisson distribution with a mean of 7.5 per hour. We call the number of arrivals x and the waiting time for a taxi to come W.

a) What is the mean number of arrivals per half hour? Per minute?
   3.75, 0.125

b) What is the probability of having to wait more than 10 minutes for a taxi?
   $P(W > 10) = P(0\ taxis\ in\ 10\ minutes)$
   $= P(x = 0) = \frac{\lambda^0}{0!} e^{-\lambda}$
   $= e^{-\lambda} = e^{-\frac{1}{8} \times 10}$
   $= 0.2865$

Let us note that if we measure time in minutes, the mean number of taxis per unit time.
The CDF for W is
$P(W \le t) = 1 - P(W > t)$
$= 1 - e^{-\frac{1}{8}t} = 1 - e^{-\lambda t}$

Finally, CDF:
$F(t) = \begin{cases} 1 - e^{-\lambda t}; if\ t > 0 \\ 0; if\ t < 0 \end{cases}$

The density function for W is
$F'(t) = (1 - e^{-\lambda t})'$
$F'(t) = 0 - e^{-\lambda t}(-\lambda) = \lambda e^{-\lambda t}$ <-- We will never use this formula in ECON 227

Insert:

$E(x) = \mu_x = \mu(x) = \sum x_i P(x_i)$

For continuous W
$E(W) = \int_{-\infty}^{\infty} t f(t) dt$

Again, not used in ECON 227.

Exercise

For W with $f(t) = \lambda e^{-\lambda t}$
$E(x) = mean\ waiting\ time$
$= \int_0^{\infty} t\lambda e^{-\lambda t} dt = \frac{1}{\lambda}$

$E(W) = \frac{1}{\lambda} \therefore \lambda = \frac{1}{E(W)}$
Do not forget this!!

Example

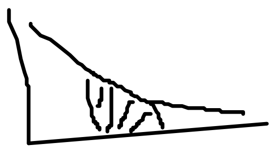
Question: Suppose the fire alarms at a fire station arrive according to an exponential distribution with a mean waiting time of 20 minutes.

a) What is the probability of having to wait more than 30 minutes for an alarm?
b) What is the probability that the next alarm will happen between 15 and 35 minutes from now?
c) What is the 55th percentile of alarm waiting times

Solution:

a)
$\lambda = \frac{1}{20}$
$P(W > 30) = e^{-\lambda t} = e^{-\frac{1}{20}(30)} = 0.223$

b)
$= area\ to\ the\ right\ of\ 15, minus\ area\ to\ the\ right\ of\ 35$
$= e^{-\frac{1}{20}(15)} - e^{-\frac{1}{20}(35)}$
$= 0.472 - 0.174 \approx 0.298$



c)
$1 - e^{-\lambda t} = 0.55$
$0.45 = e^{-\frac{1}{20}t}$
$\ln(0.45) = -\frac{1}{20}t$
$t = -20ln0.45 = 15.97\ min \approx 16min$

Lecture 13/11/04

November 4, 2013
2:34 PM

A Monday Spent with the Exponential

$F(t) = 1 - e^{-\lambda t}$
$f'(t) = \lambda e^{-\lambda t}$

Given a density function,

$area = CDF = f(t) = P(W \leq t)$

$E(W) = \frac{1}{\lambda}$
$\sigma(W) = \frac{1}{\lambda}$

$\lambda = \frac{1}{E(W)}$

But this is the same value of λ in the relationship where the expected value of W for the poisson random variable

Find the mistake:
Professor X says that the waiting times between arrivals at the Roddick Gates are exponentially distributed with a mean of 8 minutes. How would his students answer the following questions.

a) A student has just missed the bus. What is the probability of having to wait more than 10 minutes for the next one?

   Random variable is W, $\lambda = \frac{1}{8}$ lambda is 1 over the expected wait time, the unit is per minute

   Answer: $P(> 10) = e^{-\lambda t} = e^{-\frac{1}{8}(10)} = 0.2865$

b) What is the 75th percentile of waiting times?

   $P(W \leq t) = 0.75$
   $1 - e^{-\lambda t} = 0.75$
   $e^{-\lambda t} = 0.25$
   $-\lambda t = \ln 0.25$
   $-\frac{1}{8}t = \ln 0.25$
   $t = 11.088\ minutes$

c) The previous bus left 5 minutes ago. What is the probability that you will have to wait more than a further 10 minutes for the next bus?

   $P(W > 15|W > 5)\ ; P(A|B) = \frac{P(A \cap B)}{P(B)}$
   $= \frac{P((W > 15) \cap (W > 5))}{P(W > 5)}$
   $= \frac{P(W > 15)}{P(W > 5)} = \frac{e^{-\lambda(15)}}{e^{-\lambda(5)}} = 0.2865$
   Mistake found! The bus should not be exponentially distributed because busses are not equally likely to arrive throughout time because they follow a schedule

The phenomenon in part C is called the no memory property of the exponential distribution. Probabilities in the future are independent of what happened in the past.

d) What proportion of waiting times are expected to be within 2 standard deviations of the mean (lambda is unknown)

$$mean + 2stdev = \frac{1}{\lambda} + 2\frac{1}{\lambda} = -\frac{1}{\lambda}$$
$$mean - 2stdev = \frac{1}{\lambda} - 2\frac{1}{\lambda} = 3\frac{1}{\lambda}$$
$$P\left(-\frac{1}{\lambda} \leq W \leq 3\frac{1}{\lambda}\right) = 1 - e^{-\lambda t} = 1 - e^{-\lambda\left(\frac{3}{\lambda}\right)} = 0.9502$$

Lecture 13/11/06

November 6, 2013
2:40 PM

Things to watch out for:

1. Poisson and exponential - Poisson is discrete and exponential is continuous. Peanuts are discrete peanut butter is continuous.
2. $\lambda = \frac{1}{E(W)}; exponential\ \lambda = E(x)\ ; Poisson$
3. λ is the same number for the Poisson and exponential twins, nevertheless they have to be adjusted for a change in units/scale. 1/300 seconds is actually 2 / ten minutes.

We were supposed to do the bell-shaped curve, but instead we are doing…

Correlation

Two random variables, x & y are said to be positively correlated if large values of x tend to be associated with large values of y and if small values of x tend to be associated with small values of y.

Two random variables, x & y are said to be positively correlated if large values of x tend to be associated with small values of y and if small values of x tend to be associated with large values of y.

Covariance and correlation coefficient

These quantities measure correlation. Initially, at least, we shall approach these numbers via a joint-probability table.

y\x	5	12	25	
0	0.11	0.07	0.20	0.38
3	0.10	0.09	0.09	0.28
10	0.23	0.03	0.08	0.34
	0.44	0.19	0.37	1.00

The co-variance of x and y is written as $Cov(x, y) = \sigma_{xy}$

$$= \sum (x_i - E(x))(y_i - E(y))P(x_i, y_i)$$

$E(x) = 13.73; \sigma(x) = 9.005392829$
$E(y) = 4.24; \sigma(y) = 4.306088713$

A bunch of addition using the summation above later… $Cov(x, y) = -11.6252$

$$Correlation\ coefficient\ "r" = \frac{Cov(x,y)}{\sigma(x)\sigma(y)} = -\frac{11.6252}{(9.005)(4.3061)} = -0.2998$$

Lecture 13/11/11

November 11, 2013
2:42 PM

Continuing covariance

y\x	2	6	10	
-3	0.10	0.15	0.06	0.31
4	0.28	0.31	0.10	0.69
	0.38	0.46	0.16	1.00

$$Cov(x,y) = \sum(x_i - E(x))(y_i - E(x))P(x_i, y_i)$$
$$E(x) = 5.12$$
$$\sigma(x) = 2.804567703$$
$$E(y) = 1.83$$
$$\sigma(y) = 3.237452702$$

$$Cov(x,y)$$
$$= (2-5.12)(-3-1.83)(0.10) + (2-5.12)(4-1.83)(0.28) + (6-5.12)(-3-1.83)(0.15)$$
$$+ (6-5.12)(4-1.83)(0.31) + (10-5.12)(-3-1.83)(0.06) + (10-5.12)(4-1.83)(0.10) = -0.7896$$

$$Cov(x,y) = r\sigma_x\sigma_y$$

The correlation coefficient (rho)

$$P(x,y) = P_{xy} = \sigma_{xy} = \frac{Cov(x,y)}{\sigma(x)\sigma(y)}$$

On the calculator, r gives rho if there are frequencies or probabilities. The purpose of the denominator in rho is to keep the value between -1 and +1.

Our rho is -0.086963653 is too small to justify being a strong correlation, but the relationship is nonetheless negative

Some Technical Formulas

$$\sigma^2(x+y) = \sigma^2(x) + \sigma^2(y) + 2Cov(x,y)$$
$$\sigma^2(x-y) = \sigma^2(x) + \sigma^2(y) - 2Cov(x,y)$$

(from binomial expansion!)

If x and y are two discrete random variables, then x+y is a new random variable for which the values are all the possible sums of possible x and possible y values.

$$Let\ T = x + y$$

x+y	T	P(x+y)
2+(-3)	-1	0.10
6+(-3)	3	0.15
10+(-3)	7	0.06
2+4	6	0.28
6+4	10	0.31
10+4	14	0.10

$$\sigma^2(T) = \sigma^2(x+y)$$
$$= (4.094813793)^2$$
$$= 16.7675$$

$$\sigma^2(x+y) = \sigma^2(x) + \sigma^2(y) + 2Cov(x,y)$$
$$= 7.8656 + 10.481172 + 2(-0.7896)$$
$$= 16.7675$$

Wow such right much correct wow
wow

$$Let\ Q = x - y$$

(repeat the above exercise for x-y and you should get $\sigma^2(x-y) = 19.9259$)

If x & y are independent random variables, then $Cov(x, y) = 0$ ∴ for an independent random variable…
$\sigma^2(x+y) = \sigma^2(x) + \sigma^2(y)$
$\sigma^2(x-y) = \sigma^2(x) + \sigma^2(y)$

Lecture 13/11/13

November 13, 2013
2:40 PM

From last time

$Cov(x, y) = \sum(x_i - E(x))(y_i - E(x))P(x_i, y_i)$

If x & y are independent random variables, then $Cov(x, y) = 0 \therefore$ for an independent random variable…
$\sigma^2(x + y) = \sigma^2(x) + \sigma^2(y)$
$\sigma^2(x - y) = \sigma^2(x) + \sigma^2(y)$

Since the $Cov(x, y)$ is zero for statistically independent random variables.

Example of indirect computation of Rho

In the Eurelian gromment industry, employees are always hired as couples. The salaries of husbands and wives are perhaps independent, perhaps not.

We obtain the following summarized data:

$E(H) = 21600$
$E(W) = 19800$
$\sigma(H) = 2800$
$\sigma(W) = 3200$

$F = family\ income = H + W$
$E(F) = 41400$
$\sigma(F) = 3600$

Determine whether the income of husbands and wives are correlated.

Solution

$\sigma^2(x + y) = \sigma^2(x) + \sigma^2(y) + Cov(x, y)$
$3600^2 = 2800^2 + 3200^2 + 2Cov(H, W)$
$-5120000 = 2Cov(H, W)$
$Cov(H, W) = -2560000$

$r = \dfrac{Cov(x, y)}{\sigma(x)\sigma(y)}$
$r = -\dfrac{2560000}{(2800)(3200)}$
$r = -0.2857$

Application to Sampling

A simple random sample of size n is a subset of the population selected in such a way that every possible subset of size n is equally likely to be chosen.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Let us stipulate that the randomness makes the different x variables independent of each other.

Background calculation

$\sigma^2(x_1 + x_2 + x_3 + \cdots + x_n)$
$= \sigma^2(x_1) + \sigma^2(x_2) + \sigma^2(x_3) + \cdots + \sigma^2(x_n)$
$+ \textit{whole bunch of } Cov(\ldots) \textit{ terms, but these are } 0 \textit{ because of randomness}$

This is why statisticians wants to choose statistically random samples

$= \sigma^2 + \sigma^2 + \sigma^2 + \cdots + \sigma^2 = n\sigma^2$
$\sigma^2(x_1 + x_2 + x_3 + \cdots + x_n) = n\sigma^2$

$\sigma(\sum x_i) = \sqrt{n\sigma^2} = \sqrt{n}\sigma$

$\bar{x} = \dfrac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$

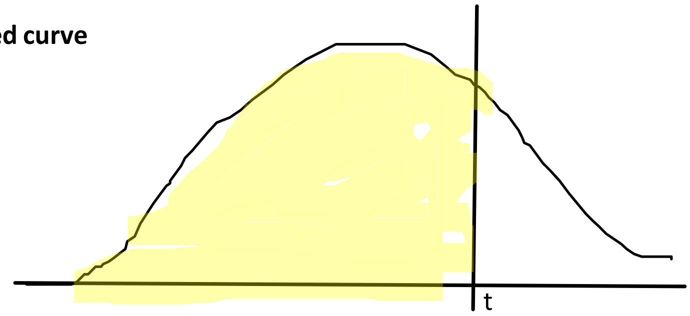
$= \dfrac{\sum x_i}{n}$

$= \sigma(\bar{x}) = \sigma\left(\dfrac{1}{n}\sum x_i\right)$

$= \dfrac{1}{n}\sqrt{n}\sigma$

$= \dfrac{\sigma}{\sqrt{n}}$

The bell-shaped curve



This is the density curve for a continuous random variable called the standard normal random variable Z.

The area $= P(Z \leq t)$

The numbers listed in the Z tables are the values of the CDF $F(t) = P(Z \leq t) = \textit{area to the left of } t$

e.g. $P(Z \leq 1.56) = 0.9406$

In this case t = 1.56

Lecture 13/11/18

November 18, 2013
2:44 PM

Z-tables

$P(a \leq 0 \leq b) = area\ to\ the\ left\ of\ b\ minus\ the\ area\ to\ the\ left\ of\ a$

Find the following percentiles of Z

Use the tables backwards

95th	1.645
90th	1.28
99th	2.33
68th	0.47

General Normal Random Variables

Let X be a continuous random variable with a mean μ, standard deviation σ. X is said to be normally distributed if the transformed random variable

$$\frac{X-\mu}{\sigma}\ (ie. z - score)$$

has the same distribution as Z.

In a math course (so not in ECON 227)

A continuous random variable with mean μ, standard deviation σ is called normally distributed if its density function $f(x) = \frac{1}{(\sqrt{2\pi})^{\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

</not in ECON 227>

Example: At a fish farm, trout are considered marketable at the age of 8 months. Marketable trout have weights that are (approximately) normally distributed with mean 820 grams stdev 122 grams.

a) What proportion of the trout weigh between 700 and 900 grams?
b) A worker at the farm scoops up 80 of the fish with a net, about how many of the scooped fish weigh more than 600 grams?
c) What is the 95th percentile of trout weights?

Solutions:

a)

$$P(700 \leq X \leq 900) = P\left(\frac{700-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{900-\mu}{\sigma}\right)$$
$$= P\left(\frac{700-820}{122} \leq Z \leq \frac{900-820}{122}\right)$$
$$= P(-0.99 \leq Z \leq 0.66)$$
$$= entry\ for\ 0.66 - entry\ for\ -0.99$$
$$= 0.7454 - 0.1611$$
$$= 0.5833$$

b)

$$P(X > 600) = P\left(\frac{X-\mu}{\sigma} > \frac{600-\mu}{\sigma}\right)$$
$$= P\left(Z > \frac{600-820}{122}\right)$$
$$= P(Z > -1.80)$$
$$= 1 - area\ left\ of\ -1.80\ OR = area\ left\ of\ 1.80$$
$$= 1 - 0.0359\ OR = 0.9641$$
$$= 0.9641$$

$$0.9641 \times 80\ fish \approx 77\ fish$$

c)

Reverse lookup --> Z=1.645

$$1.645 = \frac{X-\mu}{\sigma}$$
$$X = \mu + Z\sigma$$
$$X = 1021\ grams$$

Recall:

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}} = std\ error\ of\ \overline{x}$$

A standard error is the standard deviation of a statistic.

$$\frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}\ is\ a\ z-score\ for\ \overline{x}$$

Theoremlet

Let X be normally distributed with mean μ, standard deviation σ. Let $\overline{x}$ be the random variable for which the numerical value is the mean of a randomly selected sample of size n. Then $\overline{x}$ is also normally distributed (mean μ, standard deviation $\frac{\sigma}{\sqrt{n}} = standard\ error$) and the transformed random variable $\frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ has the same distribution as Z.

d) If the fish in the sample from part b) have a random sample of weights, what is the probability that the mean weight in the sample is less than 850grams?

Solution:

$$P(\overline{x} < 850)$$
$$= P\left(\frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < \frac{850-\mu}{\frac{\sigma}{\sqrt{n}}}\right)$$
$$= P\left(Z < \frac{850-820}{\frac{122}{\sqrt{80}}}\right)$$
$$= P(Z < 2.20)$$
$$= 0.9861$$

Review 13/11/20

November 20, 2013
2:39 PM

Normal Distribution

We have these two:

$$\frac{X-\mu}{\sigma}\ and\ \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

Both are for z-scores but left one is for probability of x and on the right is the one for the probability of x-bar (average value of all x, so the keyword is *mean*).

A Paraphrase of a Former Exam Question

The owner and cook in a short order restaurant have both taken ECON 227. The veggie quickie breakfast is under discussion. The owner and cook agree that it takes an average of about 3 minutes to prepare. The owner is of the opinion that the lengths of time taken to prepare the breakfast have an exponential distribution whereas the cook feels the normal distribution is appropriate with standard deviation = 1.5.

a) What is the probability that a VQ breakfast will take longer than 4 minutes to prepare?
b) In a randomly-selected 10 minute period, what is the probability that the cook will finish at least 3 VQ if they are prepared one after another.
c) What is the 90th percentile of preparation times for the VQ?

Owner's solutions (exponential distribution)

a) $P(W>4)=e^{-\lambda t}$
$\lambda=\frac{1}{E(W)}=\frac{1\,VQ}{3\,minute}$
$P(W>4)=e^{\frac{1}{3}(4)}=0.2636$
b) $P(X\geq 3)=1-P(0)-P(1)-P(2)$
$\lambda=\frac{1\,VQ}{3\,minute}\ again!$
$=\frac{1\,VQ}{3\,minute}\times 10\,minute$
$P(X\geq 3)=1-\frac{\left(3\frac{1}{3}\right)^0}{0!}-\frac{\left(3\frac{1}{3}\right)^1}{1!}-\frac{\left(3\frac{1}{3}\right)^2}{2!}$
$=0.647$
c) $0.9=1-e^{-\lambda t}$
$e^{-\lambda t}=0.1$
$-\lambda t=\ln 0.1$
$-\frac{1}{3}t=2.30258$
$t\cong 6.9\,minutes$

Cook's solutions (normal distribution)

a) $P(X>4)=P\left(\frac{X-\mu}{\sigma}>\frac{4-\mu}{\sigma}\right)$
$P\left(Z>\frac{4-3}{1.5}\right)=P(Z>0.67)$
$from\ tables\ \rightarrow P=1-0.7486=0.2514$
b) Only way to finish 3 in ten minutes is to finish three with an a maximum average of 10/3

minutes, so this question is about means and therefore we use x-bar.

$$P\left(\bar{X} \leq 3\frac{1}{3}\right)$$
$$= P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{3\frac{1}{3}-3}{\frac{1.5}{\sqrt{3}}}\right)$$
$$= P(Z \leq 0.38)$$
$$= 0.648$$

c) $90th\ percentile, from\ tables \rightarrow z = 1.28$
$$X = \mu + z\sigma$$
$$X = 3 + 1.28(1.5)$$
$$X \cong 4.9\ minutes$$

Semester one exam material ends here

A Peek at 2014

Central Limit Theorem

Suppose we have an infinite population (works well for very large populations too) with mean μ, standard deviation σ. Suppose a random sample of size n is selected; and $\bar{x}$ is the random variable for which the numerical value is the sample mean. If n is large (how large?) then even if the population is not normally distributed $\bar{x}$ is approximated distributed and both $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ (rarely used) and $\frac{\bar{X}-\mu}{\frac{s_X}{\sqrt{n}}}$ (very frequently used) have approximately the same distribution as Z.

This theorem is what makes statistics work.

Example

In Eurelia, in the adult population, 15% wear glasses. Every Eurelian adult is coded by the government as 1 if the person wears glasses and 0 if not. This gives us a large population with only 1s and 0s. Find μ & σ of this population.

x	P(x)
0	0.85
1	0.15

$$E(x) = mean = 0.15$$
$$= proportion\ p\ of\ glasses\ wearers$$
$$so \ldots \mu = p$$

$$\sigma = 0.35707142$$

$$\sqrt{p(1-p)}$$
$$= 0.35707142$$
$$so\ \ldots \sigma = \sqrt{p(1-p)}$$

Review 13/11/25

November 25, 2013
2:38 PM

- Exam may include tree diagrams, but not EDA (so no pre-quiz material)

Q1 material not on final

2. **An analysis of a large number of speeding tickets includes the following joint-probability table.**

		Posted Limit KM/h				
		30	**50**	**90**	**100**	
	$75	**0.06**	**0.10**	**0.07**	**0.08**	0.31
Fine:	**$125**	**0.06**	**0.12**	**0.18**	**0.04**	0.40
	$250	**0.12**	**0.08**	**0.05**	**0.04**	0.29
		0.24	0.30	0.30	0.16	

a) **Which posted limit has the largest mean fine?** Calculator question… ANS: 30

b) **What is the overall standard deviation in the fines?**

Use marginal values… ANS:69.83

c) **If the posted limit is 50 KM/h what is the probability that the fine is different from $125?**

Conditional probability… ANS: 0.60

d) **Let A be the event that the posted limit is 50 KM/h, and B be the event that the fine is $125. Determine whether A and B are statistically independent .**

$P(A \cap B) = P(A)P(B)$… statistically independent

e) Find p & the covariance for this table. Calculator question… ANS: Covariance = -395.9

3. **In Eurelia sixty percent of the population are baby boomers and fifteen percent are golden agers . Twenty-eight percent of golden-agers and twelve percent of baby-boomers get flu shots. In the rest of the population (those who are neither baby-boomers nor golden-agers) only one percent get flu shots.** Tree diagram

a) **What proportion of those who get flu shots are baby-boomers?**

Conditional probability… $P(BB|F) = \frac{P(BB|F)}{P(F)}$

$= \frac{0.072}{0.072 + 0.042 + 0.0025}$

$= \frac{0.072}{0.1165} = 0.618$

b) **If 1000 Eurelians are selected at random, about how many of those selected will be either golden-agers or one of those who do not get flu shots?**

$0.9255 \times 1000 = 925.5$
So 925 or 926

Conditions for hypergeometric
1. A finite number n of things
2. N=N1 successes + N2 failures
3. A random sample of size n is selected

4. **An accounting firm has been asked to audit a government grants programme . One of the projects has forty documents, of which ten have records of inappropriate transactions.**

a) **If a random sample of ten of the documents is selected, what is the probability that the sample will include more than two of the documents with records of inappropriate transactions?**

$1 - P(0) - P(1) - P(2)$
$= 1 - \frac{\binom{10}{0}\binom{30}{10}}{\binom{40}{10}} - \frac{\binom{10}{1}\binom{30}{9}}{\binom{40}{10}} - \frac{\binom{10}{2}\binom{30}{8}}{\binom{40}{10}}$
$= 0.485$

b) **What is the smallest sample size that will give a probability of more than 90% of including at least one of the documents with records of inappropriate transactions?**

Trial and error
$1 - \frac{\binom{10}{0}\binom{30}{n}}{\binom{40}{n}} > 0.90$
$= \frac{\binom{30}{n}}{\binom{40}{n}} < 0.1$
$n = 8\ (0.076)$

5. **A.J.Jones works as an inspector of airplane engines . The proportion of engines with misaligned combobulators is twelve percent . For the purposes of this question you may suppose that this is the probability of misalignment for every engine that comes under Jones's scrutiny .**

Binomial distribution

a) **If Jones inspects 200 engines, what is the expected number of misaligned combobulators amongst them? What is the standard deviation in the number of misaligned combobulators amongst them?**

$E(x) = np = 24$
$\sigma(x) = \sqrt{npq} \cong 4.6$

b) **What is the probability of at most 4 misaligned combobulators amongst the first 50 engines?**

$P(0) + P(1) + P(2) + P(3) + P(4)$
$= \binom{50}{0}(0.12)^0(0.88)^{50} + \cdots + \binom{50}{4}(0.12)^4(0.88)^{46}$
$= 0.268$

c) **What is the median number of misaligned combobulators amongst the first 50 engines?**

Keep adding from b)
$P(5) = 0.167$... cumulative 0.435
$P(6) = 0.171$ ... cumulative 0.606, total crosses 0.5

$\lambda = \frac{1\ inspections}{240\ seconds}$

6. **The lengths of time that A.J.Jones requires to inspect engines have an exponential distribution with a mean of 240 seconds.**

a) **If Jones inspects 200 engines, about how many of them will take him longer than 300 seconds to inspect?**

$P(W > 300) = e^{-\lambda t}$
$= e^{-\frac{1}{240}(300)}$
$= 0.2865$
ANS: 0.2865 x 200 ~= 57

b) **What is the 95th percentile of Jones's inspection times?**

$e^{-\lambda t} = 0.05$
$e^{-\frac{1}{240}t} = 0.05$
$t = -240 \ln 0.05$
$t = 119\ seconds$

c) **What is the probability that Jones will complete the inspection of more than three engines in a randomly-selected twenty-minute period?**

$\lambda = \frac{1\ inspections}{240\ seconds}(1200\ seconds) = 5\frac{inspections}{20\ minutes}$
$1 - P(0) - P(1) - P(2) - P(3)$
$= 1 - \frac{5^0}{0!}e^{-5} - \frac{5^1}{1!}e^{-5} - \frac{5^2}{2!}e^{-5} - \frac{5^3}{3!}e^{-5} \cong 0.735$

7. **The lengths of time that J.R.Smith requires to inspect engines have a normal distribution with a mean of 240 seconds, standard deviation 80 seconds .**

$P(X > 300)$
$= P\left(Z > \frac{300 - 240}{80}\right)$
$= P(Z > 0.75)$
$= 1 - 0.7734$
$= 0.2266$

a) **If Smith inspects 200 engines, about how many of them will take her longer than 300 seconds to inspect?**

b) **What is the 95th percentile of Smith's inspection times?**

Reverse lookup... 95th percentile of Z is 1.645
$X = \mu + Z\sigma = 240 + (1.645)(80) = 371.6$

c) **For the 200 engines that Smith inspects, what is the probability that her mean inspection time will be less than 250 seconds ?**

$P(x < 250) = P\left(Z < \frac{250 - 240}{\frac{80}{\sqrt{200}}}\right)$

$P(Z < 1.77) \cong 0.96$

d) **Who has the smaller standard deviation in inspection times: A.J.Jones or J.R.Smith ?**

Exponential, mean = STDEV

Review 13/11/27

November 27, 2013
2:42 PM

1. a) **Consider the following data set in stem-and-leaf notation.**

3	**09**
4	**238**
5	**5789**
6	
7	**09**

No stem and leaf on final

Determine whether there are any outliers using the z-score criterion.

b) **For the joint probability distribution**

		X		
		- 2	**0**	**4**
Y	**0**	**0.1**	**0.2**	**0.1**
	5	**0.2**	**0.1**	**0.3**

let T = X + Y. Find the probabilities in the PDF of T:

t_i	**- 2**	**0**	**3**	**4**	**5**	**9**
$P(t_i)$	**0.xx**	**0.xx**	**0.xx**	**0.xx**	**0.xx**	**0.xx**
	0.1	0.2	0.2	0.1	0.1	0.3

c) **For the distribution in b) verify numerically that**

$$\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y) + 2Cov(X,Y).$$

$\sigma^2(x+y) = \sigma^2(T)$
$= (3.8209946)^2$
$= 14.6$

$\sigma_x = 2.9690465$
$\sigma_x^2 = 6.6$
$\sigma_y = 2.4494897$
$\sigma_y^2 = 6$

$r = 0.1589$
$Cov(X,Y) = r\sigma_x\sigma_y$
$Cov(X,Y) = 1$

$\sigma(x)^2 + \sigma(y)^2 + 2Cov(X,Y) = 14.6$
Checks out!

2. **East Eurelian Airways is a small carrier that only has three destinations. Thirty percent of flights go to Apica, fifty percent go to Begonia, and the remaining twenty percent go to Calendula. Half of the Apica flights are return flights, one-third of Begonia flights are return flights, and one-fourth of Calendula flights are return flights.**

a) **What proportion of return flights have Apica as the destination?**

Tree diagram Conditional probability because a reference is made to a subset… ANS~= 41%

A given event is on the right side of the line (e.g. AJ Jones has purchased a one way ticked, what is the probability of him travelling to Apica.)

Taking a proportion of a subset - conditional probability…

$$P(B|R) = \frac{P(B \cap R)}{P(R)} = \frac{0.335}{0.6333} \cong 53\%$$

b) **What proportion of the one-way flights have Begonia as the destination?**

c) **Let A be the event that a flight has Apica as its destination, and R be the event that the flight is a return flight. Determine whether A and R are statistically independent.**

$P(A|R) = 0.15$
$P(A)P(R) = 0.3 \times 0.365 = 0.11$
$Since\ P(A \cap R) \neq P(A)P(R)$
A and R are dependent events

3. **Acme, Incorporated, has tens of thousands of employees world-wide. A joint-probability table has been prepared based on employee data.**

Taking a proportion of a subset - conditional probability...

$$P(B|R) = \frac{P(B \cap R)}{P(R)} = \frac{0.335}{0.6333} \cong 53\%$$

b) What proportion of the one-way flights have Begonia as the destination?

c) Let A be the event that a flight has Apica as its destination, and R be the event that the flight is a return flight. Determine whether A and R are statistically independent.

$P(A|R) = 0.15$
$P(A)P(R) = 0.3 \times 0.365 = 0.11$
$Since\ P(A \cap R) \neq P(A)P(R)$
A and R are dependent events

3. Acme, Incorporated, has tens of thousands of employees world-wide. A joint-probability table has been prepared based on employee data.

	Female	Male
University Degree	0.4	0.3
No Degree	0.2	0.1

Grey area... strictly speaking this is hyper-geometric but binomial still works well because taking away a person from tens of thousands doesn't change the probabilities by much.

Every week six employees are selected at random to give employee feedback.

a) What is the probability that at least three of the six selected will be women with university degrees?

b) What is the average number of university graduates in the samples? What is the standard deviation in the numbers of university graduates in the samples?

c) Once a year a special sample of 10 women employees is selected. What is the probability that at least nine of the women in the sample have university degrees?

a)
Hyper geometric take N=10000
4000 are successes 6000 are failures

$$= \frac{\binom{4000}{3}\binom{6000}{3}}{\binom{10000}{6}} + \frac{\binom{4000}{4}\binom{6000}{2}}{\binom{10000}{6}} + \frac{\binom{4000}{5}\binom{6000}{1}}{\binom{10000}{6}} + \frac{\binom{4000}{6}\binom{6000}{0}}{\binom{10000}{6}}$$

$= 0.45568$

Binomial

$$= \binom{6}{3}(0.4)^3(0.6)^3 + \binom{6}{4}(0.4)^4(0.6)^2 + \binom{6}{5}(0.4)^5(0.6)^1 + \binom{6}{6}(0.4)^6(0.6)^0$$

$= 0.45568$

b)
$np = 6(0.2) = 4.2 = E(x)$
$\sqrt{npq} = \sqrt{6(0.7)(0.3)} = 1.12249$

c) Conditional probability (probability of university degree given it's a woman)

$$\binom{10}{9}(0.6667)^9(0.3333)^1 + \binom{10}{10}(0.6667)^{10}(0.3333)^0 = 0.104$$

4. The fifty employees at a branch office of Acme, Incorporated, have been cross-categorized as shown. Similar to last question but **not** grey area

	Female	Male
University Degree	20	15
No Degree	10	5

$$1 - P(0) - P(1) = 1 - \frac{\binom{20}{0}\binom{30}{6}}{\binom{50}{6}} - \frac{\binom{20}{1}\binom{30}{5}}{\binom{50}{6}}$$

Six of the employees are selected at random.

a) What is the probability that the sample includes at least two men?

b) What is the probability that the sample includes at most two

a) **What is the probability that the sample includes at least two men?**

b) **What is the probability that the sample includes at most two employees with university education?** $= P(0) + P(1) + P(2) = 0.058$

c) **What is the probability that the sample includes at least one man and at least one employee with a university degree?**

$$= 1 - \frac{\binom{30}{6}\binom{20}{0}}{\binom{50}{6}} |women| - \frac{\binom{15}{6}\binom{35}{0}}{\binom{50}{6}} |no\ degrees| + \frac{\binom{10}{6}\binom{40}{0}}{\binom{50}{6}} |women\ without\ degrees\ got\ taken\ out\ twice| = 0.962$$

5. At Acme, Incorporated, it is claimed that the probability of an accident during a work day is eight-tenths of one percent. There are two hundred and twenty-five work days in a year.

a) **What are the mean and standard deviation in the number of accidents per year?** See myCourses - binomial pack practice problems, answers in Q11

b) **What is the probability of more than the mean number of accidents in a year?**

c) **What is the probability of having to wait until after the first hundred work days in the year before the first accident occurs?**

6. A time-and-motion-study consultant has been hired at Eurelia Industries Limited. She has identified a certain work station as a bottleneck in production. Initial data suggest that the times required for processing pieces at this station may be treated as having an exponential distribution with a mean of three hundred seconds.

a) **If fifty pieces are processed, about how many of them take between two hundred forty and three hundred sixty seconds to process?**

b) **What is the probability that more than the mean number of pieces will be processed in a ten-minute period?**

c) **What is the median processing time at the station?**

7. The fire department in Eurelia City holds fire drills in downtown office buildings. Records show that the times required to evacuate ten-storey buildings during fire drills are approximately normally distributed with a mean of nine minutes and a standard deviation of two minutes. In a one-month period sixty randomly-selected ten-storey buildings held fire drills.

a) **About how many of the buildings were evacuated in less than eight minutes that month?**

b) **What is the probability that the average evacuation time was between eight minutes forty-five seconds (8.75 minutes) and nine minutes fifteen seconds (9.25 minutes)?**

c) **The fire department wishes to specify a length of time such that ninety-nine percent of ten-storey buildings will be evacuated in that amount of time or less. What length of time should be specified?**

a)

$P(x < 8) = P(Z < \frac{8-9}{\ldots})$

that ninety-nine percent of ten-storey buildings will be evacuated in that amount of time or less. What length of time should be specified?

a)

$P(x < 8) = P\left(Z < \frac{8-9}{2}\right)$

$= P(Z < -0.5) = -0.3085$

ANS: -0.3085 x 60 = -18.51

b)

$P(8.75 < X < 9.25) = P\left(\frac{8.75-9}{\frac{2}{\sqrt{60}}} < Z < \frac{9.25-9}{\frac{2}{\sqrt{60}}}\right)$

$P(-0.96 < z < 0.96) = 0.8315 - 0.1685 = 0.663$

c)

99th percentile of Z = 2.33

$X = \mu + z\sigma$

$= 9 + 2.33(2)$

$= 13.66\ minues$

Review 13/12/03

December 3, 2013
2:38 PM

When to use $\sqrt{\lambda}$ or $\sqrt{npq}$? Difference of poisson vs. exponential.

In accidents question, $\sqrt{npq} = \sqrt{225 \times 0.008 \times 0.992} = 1.336$

$\sqrt{\lambda} = \sqrt{1.8} = 1.342$

So not too far, but stick to what you chose when you started the problem.

4. **At Eurelian Engineering, Incorporated, plans are in progress to institute statistical quality control. The newly-appointed head of the quality-control department is interested in establishing benchmarks. Here are some questions typical of those she has asked the team to answer.**

a) **If there are 3 defective components in a randomly-selected sample of 60 components, what is the probability that at least one of the 3 defectives will appear amongst the first 10 selected?**

b) **What is the probability that the first three components in the sample of size 60 were a defective, followed by a non-defective, followed by another non-defective?**

c) **If the true overall proportion of defectives is 3%, how likely is it that 3 or more defectives would be present in a randomly-selected sample of size 60?**

a)
$P(X \geq 1)$
$= 1 - P(0)$
$$= 1 - \frac{\binom{3}{0}\binom{37}{10}}{\binom{60}{10}} = 0.427$$

b)
$$\frac{3}{60} \times \frac{57}{59} \times \frac{56}{58} = 0.0466$$

c)
You can do it as a binomial but strictly speaking it is a finite sample so you can do the hypergeometric while picking some large number to be the total number.

Hypergeometric
$$\frac{\binom{300}{3}\binom{9700}{57}}{\binom{10000}{60}} \cong 0.1631$$

Binomial
$$\binom{60}{3}(0.03)^3(0.97)^{57} = 0.162791836$$

$$1 - \binom{60}{3}(0.03)^0(0.97)^{60} - \binom{60}{3}(0.03)^1(0.97)^{59} - \binom{60}{3}(0.03)^2(0.97)^{58} = something$$

5. **The traffic lights near the home of A.J.Jones change colour at random times, so that Jones considers the number of changes in any given time period to have a Poisson distribution, with a mean of 20 changes per hour.**

a) **What is the probability of more than the mean number of changes in a nine-minute period?**

b) **What is the median number of changes in a nine-minute period?**

d) **What is the probability of having to wait more than 3 minutes for the light to change?**

a)
$$\lambda = 20 \times \frac{9}{60} = 3$$
$P(X > 3) = 1 - P(0) - P(1) - P(2) - P(3)$

b)

x	P(x)
0	0.05
1	0.15
2	0.225
3	0.225*

c)
$$\lambda = \frac{20}{60}$$
$P(W > 3) = e^{-\lambda t} = e^{-\frac{20}{60}(3)}$
$= 0.368$

5.5 **The anchovy distribution unit at BIG CRUST PIZZA puts anchovy fillets on 25 pizzas at once, in a random way so that every anchovy fillet is equally likely to end up on any one of the 25 pizzas. Suppose 3 of the 25 pizzas end up with no anchovy fillets. Estimate the total number of anchovy fillets put on the 25 pizzas by the anchovy dispersal unit.**

5.5 **The anchovy distribution unit at BIG CRUST PIZZA puts anchovy fillets on 25 pizzas at once, in a random way so that every anchovy fillet is equally likely to end up on any one of the 25 pizzas. Suppose 3 of the 25 pizzas end up with no anchovy fillets. Estimate the total number of anchovy fillets put on the 25 pizzas by the anchovy dispersal unit.**

6. **The lengths of time required to play college football games has been found to have approximately a normal distribution with a mean of 150 minutes, and standard deviation 18 minutes.**

a) **If 300 college games are played, about how many of them last more than 160 minutes?**

b) **What is the length of time with 33% of college games lasting longer than it?**

c) **If 12 college games are selected at random, what is the probability that the mean length of the games will be between 145 and 160 minutes?**

a)
$\mu = 150\ \sigma = 18$
$P(X > 160) = P\left(Z > \frac{160-150}{18}\right)$
$= P(Z > 0.56)$
$= 1 - 0.7123 = 0.2877$

b)
$67th\ percentile\ of\ Z = 0.44$
$\mu + Z\sigma = 150 + 0.44 \times 18$
$\cong 157.92 \approx 158$

c)
$P(145 < x < 160)$
$= P\left(\frac{145-150}{\frac{18}{\sqrt{12}}} < Z < \frac{160-150}{\frac{18}{\sqrt{12}}}\right)$
$= P(-0.96 < Z < 1.92)$
$= 0.9727 - 0.1685 = 0.8042$

Crib Sheet 13/12/05

December 5, 2013
2:13 PM

David Zhou at 2013-12-05 2:32 PM
Probability theory, including the expected value and standard deviation of random variables

Rules for statistical independence

i) $P(A|B) = P(A|\bar{B})$

ii) $P(A|B) = P(A)$

iii) $P(A \cap B) = P(A)P(B)$ *

iv) $P(B|A) = P(B|\bar{A})$

v) $P(B)A = P(B)$

Geometric, binomial, Poisson, and hypergeometric discrete random variables

Binomial discrete random variables
$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$

$E(x) = np$

$\sigma(x) = \sqrt{npq}$

Hypergeometric discrete random variables

20 women and 10 men are in a society. If 5 members are selected at random what is the probability that 3 of them will be women and 2 men.

$$\frac{\binom{20}{3}\binom{10}{2}}{\binom{30}{5}} \cong 0.35998484 \cong 36\%$$

Exponential and normal continuous random variables

Lecture 14/01/06

January 6, 2014
2:37 PM

- If switching sections, don't do so on Minerva, but if dropped by accident, the correction can be made at servicepoint
- If calculator is not two variable, you can get the SHARP EL-531

What we use from term 1

- $\bar{x}, \sigma_x, s_x, Cov(x,y), rho = \frac{Cov(x,y)}{\sigma_x \sigma_x}$
- Don't need to remember too much about distributions, but at least remember poisson distribution and normal distribution

Poisson

- $E(x) = \lambda, \sigma(x) = \sqrt{\lambda} = \sqrt{E(x)}$

Normal

- $\frac{x-\mu}{\sigma}, \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ if the pop is normal. These two have the same distribution as z

Standard error of a statistic

$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}} \approx \frac{s_x}{\sqrt{n}}$

Central Limit Theorem

- Suppose we have an infinite population (works pretty well with a large finite population too) with mean mu, std dev sigma
- Suppose we have an SRS (simple random sample) of size n, let x-bar be a random variable for which the numerical value is the mean of the SRS. Then, even if the original population is not normally distributed, if n is large, x-bar is approximately normally distributed. In particular the following all have approximately the same distribution as Z

$$\frac{\bar{p} - p}{\sqrt{\frac{p(1-\bar{p})}{n}}}, \quad \frac{\bar{p} - p}{\sqrt{\frac{p(1-\bar{p})}{n}}}$$

Lecture 14/01/08

January 8, 2014
2:40 PM

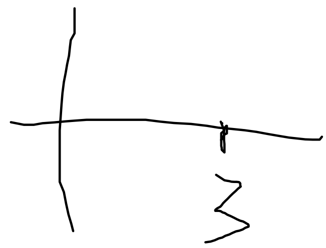
- Exam rereads with KMacK at his office: MT 4:00 - 6:00+
- KMacK says Z-tables weren't supposed to be taken

Let's begin

$$\frac{X-\mu}{\sigma}, \frac{X-\mu}{\frac{\sigma}{\sqrt{n}}}, \frac{p-p}{\sqrt{\frac{p(1-p)}{n}}}, \frac{p-p}{\sqrt{\frac{p(1-p)}{n}}}$$

Definitions

1. A parameter is a number calculated from the entire population. e.g. mu, sigma, p
2. SRS stands for simple random sample. This is a randomly selected subset of the population. If the size of the SRS is n, it must be chosen in such a way that every possible subset of size n is equally likely to be chosen.
3. A statistic is a number calculated from a sample (in ECON 227, samples will always be SRS). e.g. X, S_x, p
4. The use of a single number (statistic) to estimate a parameter is called <u>point estimation</u>.



- A problem with a point estimate is that it is highly unlikely to be exactly equal to the parameter that it is estimating. For this reason, statisticians prefer interval estimates with a margin of error.
- If the probability can be estimated that the interval estimate actually includes the true value of the parameter, then it's called a confidence interval and the probability is called the <u>confidence level</u> (or level of confidence)

Confidence interval formula

For mu, x-bar is the point estimate.

$$x \pm 1.96\frac{S_x}{\sqrt{n}}$$

margin of error

Let us suppose that n is large enough for the central limit theorem to apply

<u>Example:</u>
In Eurelia University a SRS of 100 undergraduate students had a mean GPA of 3.12, with st dev 0.11. Use the CI formula to come up with an interval estimate for mu.

$$3.12 \pm \frac{1.96(0.11)}{\sqrt{100}}$$
$$= 3.12 \pm 0.02156 \approx 3.12 \pm 0.02$$

e.g. somewhere between 3.10 and 3.14

Let us find the confidence level:

$$P(the\ CI\ actually\ includes\ true\ value\ of\ \mu)$$
$$= P\left(X - 1.96\frac{S_x}{\sqrt{n}} < \mu < X + 1.96\frac{S_x}{\sqrt{n}}\right)$$
$$= P\left(-1.96 < \frac{\mu - X}{\frac{S_x}{\sqrt{n}}} < 1.96\right)$$
$$= P\left(1.96 > \frac{X-\mu}{\frac{S_x}{\sqrt{n}}} > -1.96\right)$$
$$= P(-1.96 < Z < 1.96)$$
$$= 0.9750 - 0.0250 = 0.95$$

Pollster Jargon: Our sample of 100 Eurelian students show that the overall mean GPA has a 95% probability of being somewhere between 3.10 and 3.14.

Or even more pollster like: Our sample gives a mean GPA of 3.12. With a sample this size, the margin of error is 2 percentage points 19 times out of 20.

The 1.96 determines the confidence level.

95% CI formula for p

$$x \pm 1.96\frac{S_x}{\sqrt{n}} \to p \pm 1.96\sqrt{\frac{p(1-p)}{n-1}}$$

Example:
A pollster in Eurelia selected 1550 voters in SRS, 806 of them support the recidivist party.

Form a 95% CI for the overall RP support amongst voters

$$p = \frac{806}{1550} = 0.52$$

$$0.52 \pm 1.96\sqrt{\frac{(0.52)(0.48)}{1550}}$$

$$= 0.52 \pm 0.0249$$

$$= 0.52 \pm 0.025$$

e.g. the margin of error with a sample this size is 2.5 percentage points 19 times out of 20.

Lecture 14/01/13

January 13, 2014
2:35 PM

- KMacK getting a room ready or makeup exam
- KMacK tells a story

The adventures of Lee Grundberg

$\bar{x} \pm Z\frac{S_x}{\sqrt{n}}$

$\bar{p} \pm 1.96\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$

margin of error

1.96 is the 95% CI formula

90% CI formulas

On the Z-table, find the z-score such that the area covered is 90%... e.g. 5% on one side (find 0.05 -> 0.0505 ~ -1.64, 0.0495 ~ -1.65, average is 1.645)

99% CI formulas

Repeat steps from above... find 0.005 -> 0.0051 ~ 2.57, 0.0049 ~ 2.58, average is 2.575

Example:

A SRS of 1600 voters includes 768 supporters of AJ Jones. Let's find 90% CI, 95% CI, 99% CI for Jones' overall support.

i)

$\bar{p} = \frac{768}{1600} = 0.48$

$90\% CI\ 0.48 \pm 1.645\sqrt{\frac{(0.48)(0.52)}{1600}}$

$= 0.48 \pm 0.02$

ii)

$95\% CI\ 0.48 \pm 1.96\sqrt{\frac{(0.48)(0.52)}{1600}}$

$= 0.48 \pm 0.024$

iii)

$99\% CI\ 0.48 \pm 2.576\sqrt{\frac{(0.48)(0.52)}{1600}}$

$= 0.48 \pm 0.032$

The only way to simultaneously increase the confidence level and keep the margin of error small is to increase the sample size.

Sample Size Formulas for Proportion

$$\bar{p} \pm Z\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \bar{p} \pm E$$

We specify the confidence level usually 95% and the desired margin of error E

$$E = Z\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$
$$n = \frac{Z^2\bar{p}(1-\bar{p})}{E^2}$$

p-bar is unknown at this stage. Usually, we just use 0.5 and say that $n = \frac{Z^2(0.5)(0.5)}{E^2}$ and call it the conservative sample size formula. "Conservative" means worst-case scenario, or the highest cost sample.

If, however, you have a prior estimate p-tilde for the proportion, use it: $n = \frac{Z^2\tilde{p}(1-\tilde{p})}{E^2}$

Example:

How large a sample is needed to establish a 95% confidence interval for Jones' support if the margin of error is to be the usual 2.5%.

Two methods: conservative formula, prior estimate p-tilde = 0.4

i)
$$n = \frac{1.96^2(0.5)(0.5)}{(0.025)^2} \cong 1536.64 \approx 1537$$

ii)
$$n = \frac{1.96^2(0.4)(0.6)}{(0.025)^2} \cong 1475.5 \approx 1476$$

Lee's Adventure

January 22nd 1995 results of the Super Poll on the Quebec referendum. Usual 1537 person polls weren't enough because the difference was smaller than the margin of error. 3 groups paid 65000$ to take a bigger poll. Lee read in the Gazette that a poll of 10011 voters are taken. This many are needed in order to have a 1 percentage point margin of error (19 times out of 20).

Lee's computation:
$$n = \frac{1.96^2(0.5)(0.5)}{0.01^2} = 9604$$

Lee phoned up the newspaper then the polling company. He discovered the reason the polling company did it to have over 10000 voters polled. He discovered the polling company charged 2600$ for the poll.

Lecture 14/01/20

January 20, 2014
2:34 PM

Hypothesis testing procedure

Step 1: H_0, H_a - Null hypothesis and alternative hypothesis.

e.g. $H_0: \mu \geq 80000$ $(not\ x\ )$ $H_a: \mu < 80000$

Brandex tires -> Left tailed test

Right tailed test: $H_0: p \leq 0.25$ $H_a: p > 0.25$

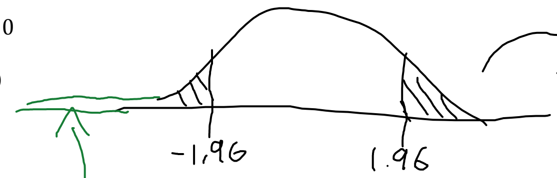
Example of a two tailed test

A filling machine is putting carbonated drinks into bottles labelled 330 mL. A random sample of 50 fillings had a mean volume of 336.4 mL, stdev 2.6mL.

We are going to test the 5 percent level of significance whether the mean fill is significantly different from 330 mL.

$H_0: \mu = 330, H_a \neq 330$

$Z(sample\ size > 30)$



$\frac{\alpha}{2} = \frac{5\%}{2} = 0.025$

Test statistic

$$= \frac{x - \mu}{\frac{s_x}{\sqrt{n}}}$$

$$= \frac{326.4 - 330}{\frac{2.6}{\sqrt{n}}} = -9.79$$

- We reject the null hypothesis.
- The mean fill is significantly different from 330 mL.
- First 3 examples were based on the central limit theorem (CLT).
- If n is large, the x-bar is approximately normal, even if the parent population is not normal.

What to do if n is small?

Question studied by William Gossett in the early 1900s. He was head brewmaster at Guinness Breweries

Fat-tail

Gossett worked out a formula for a bell shaped curve, different from the Z-curve, that works better than the Z-curve for small samples.

How do we use Student-t tables? Essentially the same way as a Z-table.

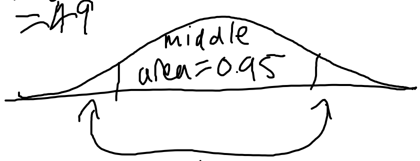
$$x \pm t\frac{s_x}{\sqrt{n}}$$

The bell shaped curve for student's t is different for every sample size. Each of the many shapes is characterized by its degrees of freedom. The degrees of freedom = n - 1.

Let's use the t tables to get a 95% CI for the mean fill of the carbonated drink.

$df = 50 - 1$
$= 49$

$$x \pm t\frac{s_x}{\sqrt{n}}$$



$$326.4 \pm \frac{2.010(2.6)}{\sqrt{n}}$$
$$326.4 \pm 0.739$$

$\frac{\alpha}{2} = 0.025$

Comparing to answer found using Z

$$326.4 \pm \frac{1.96(2.6)}{\sqrt{n}}$$
$$326.4 \pm 0.721$$

Lecture 14/01/22

January 22, 2014
2:34 PM

Student's t

Gossett (student) started his computations assuming that the data were to be sampled from a normally distributed population.

Confidence interval formula

$$\bar{x} \pm t \frac{S_x}{\sqrt{n}}$$

Degrees of freedom = n-1

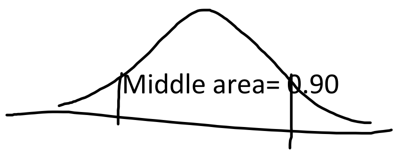
Example: In Eurelian prison inmates are served ~~iced mashed potatoes~~ ice cream every day. An incarcerated statistician has collected sample data on six randomly selected days.

Litres consumed:
231
276
119
180
244
156

Mean = 201 L
Stdev = 59.269 L

A 90% confidence interval for the mean number of litres.

Daily litres:



$$0.90 = 1 - \alpha$$
$$\alpha = 1 - 0.90 = 0.10$$

$$201 \pm 2.015 \frac{(59.269)}{\sqrt{6}}$$
$$= 201 \pm 48.756$$
$$\approx 201 \pm 49$$

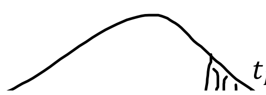
When the number is less than 30, always use T-tables, for those over 30, either Z or T is fine.

If anyone ever asks you to interpret the confidence interval, e.g. the interval above, this is what you say: Statisticians are 90% confident that the mean number of ice cream consumed daily is somewhere between 152 and 250.

Hypothesis testing question

Test at the 5% level whether the mean daily amount of ice cream being consumed in litres is significantly greater than 175 litres.

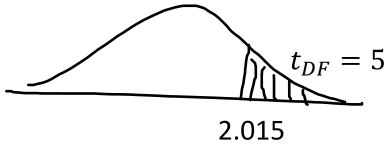
$$H_0: \mu \leq 175$$
$$H_A: \mu > 175$$



$t_{DF} = 5$

$H_0: \mu \leq 175$
$H_A: \mu > 175$

This is a right tailed test.



We did not divide alpha (0.05) by two because this is a right tailed test.

$$test\ statistic = \frac{\bar{x} - \mu}{\frac{S_x}{\sqrt{n}}} = \frac{201 - 175}{\frac{59.269}{\sqrt{6}}} = 1.07$$

We do not reject the null hypothesis H_0

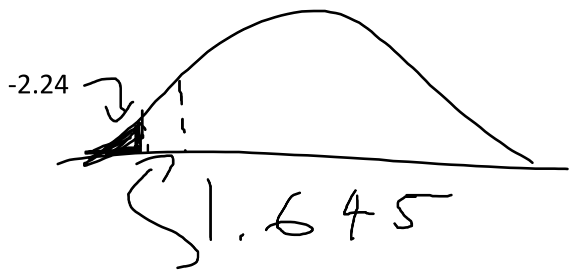
The mean consumption is not significantly greater than 175.

p-values

This became a bigger topic in statistics when computers and statistics software became commonplace.

Recall the Brandex Tire example.

$H_0: \mu \geq 80000$
$H_A: \mu < 80000$



$$test\ statistic = \frac{79780 - 80000}{\frac{1200}{\sqrt{150}}} = -2.24$$

Ghostly rejection region (rejection region doesn't apply here)

p-value definitions

1. For a left-tailed test, the p-value is the area under the density curve to the left of the test statistic.

   For the Brandex example, $p - value = P(Z < -2.25) = 0.0122 < 0.05$
   The test statistic is rejectable because it's area is less than the ghostly rejection area.
   Therefore, reject H_0. The mean tire lifetime is significantly less than 80000.

2. For a right-tailed test, the p-value is the area under the density curve to the right of the test statistic.
3. For a two-tailed test, the p-value is double the small tail area determined by the test statistic.

   Two tailed example:
   $H_0: \mu = 500$
   $H_A: \mu \neq 500$

   $$test\ statistic = \frac{540 - 500}{\frac{140}{\sqrt{49}}} = 2$$
   $= P(Z > 2)$
   $= 0.0228$
   We reject H_0 because 0.0228<0.0250 or doubling it, 0.456<0.05
   Doubling allows the direct comparison with alpha instead of alpha/2

p-value decision rule:

Reject the null hypothesis if and only if p-value < alpha (alpha is almost always 0.05). Rejecting the null hypothesis establishes statistical significance.

Lecture 14/01/27

January 27, 2014
2:29 PM

Announcements

Quiz date: 24th for this section, 17th for the morning section

Assignment: due 5th February (accepted until morning section the 7th)

Standard Error (in response to a question about the assignment)

$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}$ Usually estimated by $\frac{S_x}{\sqrt{n}}$

$$\sigma(\bar{p}) = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

A standard error is the standard deviation of a statistic. Background for two sample situations

Background for Two-Sample Situations

Recall:
$\sigma^2(x+y) = \sigma^2(x) + \sigma^2(y) + 2Cov(x,y)$
$\sigma^2(x-y) = \sigma^2(x) + \sigma^2(y) - 2Cov(x,y)$

If x and y are independent, Cov(x,y)=0. Therefore if x and y are independent,

$\sigma^2(x+y) = \sigma^2(x) + \sigma^2(y)$
$\sigma^2(x-y) = \sigma^2(x) + \sigma^2(y)0$

Apply this to x1-bar and x2-bar

$\sigma^2(\overline{x_1} - \overline{x_2}) = \sigma^2(\overline{x_1}) + \sigma^2(\overline{x_2}) - 2Cov(\overline{x_1}, \overline{x_2})$

But if $\overline{x_1}$ and $\overline{x_2}$ are idependent,
$\sigma^2(\overline{x_1} - \overline{x_2}) = \sigma^2(\overline{x_1}) + \sigma^2(\overline{x_2})$

This is accomplished by selecting the sample from the two population (e.g. women and men employees) independently of each other.

If the SRS are selected independently,
$\sigma^2(\overline{x_1} - \overline{x_2}) = \sigma^2(\overline{x_1}) + \sigma^2(\overline{x_2})$

$$\therefore \sigma^2(\overline{x_1} - \overline{x_2}) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Which is usually estimated by:

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

We say the standard error of $\overline{x_1} - \overline{x_2} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Or the standard error of the difference of means $\approx \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$

Or the standard error of the sampling distribution of the difference of means $\approx \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$

Suppose we want to test
$H_0: \mu_1 = \mu_2 \ or \ \mu_1 - \mu_2 = 0$
$H_A: \mu_1 \neq \mu_2 \ or \ \mu_1 - \mu_2 \neq 0$

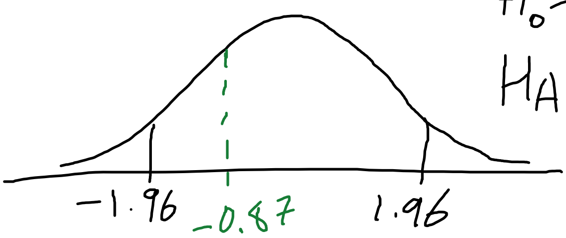
using a test statistic.

One sample test statistic $= \frac{\bar{x}-\mu_0}{\frac{s_x}{\sqrt{n}}}$
Two sample: $\overline{x_1} - \overline{x_2}$ is the test statistic that estimates $\mu_1 - \mu_2$

$$test\ statistic = \frac{(\overline{x_1} - \overline{x_2}) - 0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

We will initially treat n1, n2 as large enough that the CLT applies. In the Eurelian grommet industry:

	Women	Men
Sample size	350	250
Mean income	18983	19501
Std. dev.	6233	7812



$H_0 = \mu_1 - \mu_2 = 0$
$HA = \mu_1 - \mu_2 \neq 0$

Test statistic:

$$= \frac{(18983 - 19501)}{\sqrt{\frac{6233^2}{350}+\frac{7812^2}{250}}} = -0.87$$

We do not reject H_0, the mean incomes for women and men are not significantly different in the Eurelian grommet industry.

b) Calculate the p-value for the test.

$p - value = 2 \times P(Z < -0.87)$
$= 2 * 0.1922 = 0.3844$

Since the p-value is greater than 0.05, do not reject H_0

c) Write out a 99% CI for the difference in mean income.

$$\underbrace{\overline{x_1} - \overline{x_2}}_{\text{statistic}} \pm Z\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \rightarrow \text{standard error estimate}$$

$$18983 - 19501 \pm 2.576\sqrt{\frac{6233^2}{350}+\frac{7812^2}{250}} = -518 \pm 1535$$

Lecture 14/01/29

January 29, 2014
2:36 PM

Announcements

KmacK dresses up in a kilt.
No n given on Q4 on assignment.

Recap

$$\frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \cong Z \; if \; n_1, n_2 \; are \; large$$

What happens if n1 and n2 are small? Everyone wanted this to have a t-distribution, but it does not.

This situation is not completely solved. Current textbooks have a couple of approximations that are used with the t-distribution tables, but it's not exactly right.

1. Satterthwaite's method

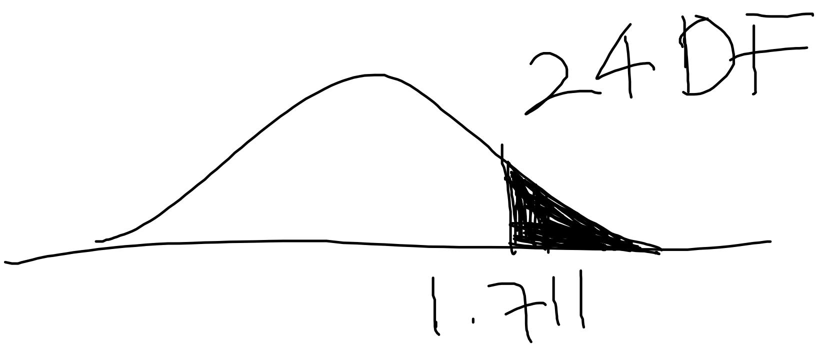
$$DF = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

   Use with t and truncate the result.
2. $DF = smaller \; of \; n_1 - 1, n_2 - 1$
   Use with t.

Example

A paleoanthropometrician has written a paper claiming that in Ancient Eurelia the women were significantly taller than the men on average. Here are the data on which the claim is based:

	Women	Men
Number of remains	12	15
Mean height	157.6 cm	149.8 cm
Std dev	8.1 cm	11.2 cm



$H_0: \mu_W \leq \mu_m$
$H_a: \mu_W > \mu_m$

$$\frac{\left(\frac{8.1^2}{12} + \frac{11.2^2}{15}\right)^2}{\frac{\left(\frac{8.1^2}{12}\right)^2}{11} + \frac{\left(\frac{11.2^2}{15}\right)^2}{14}} \cong 24.8 \approx 24$$

$$\frac{(157.6-149.8)}{\sqrt{\frac{8.1^2}{12}+\frac{11.2^2}{15}}} \cong 2.09$$

We reject the null hypothesis.

According to the paleoanthropometrician the women's mean height was significantly greater than the men's mean height

Other method

$H_0: \mu_W \leq \mu_m$
$H_a: \mu_W > \mu_m$

$DF = \min(11{,}14) = 11$

The equivalent formulas for two proportions

$\sigma^2(\bar{p}_1 - \bar{p}_2) = \sigma^2(\bar{p}_1) + \sigma^2(\bar{p}_1) - 2Cov(\bar{p}_1, \bar{p}_2)$
$\sigma^2(\bar{p}_1 - \bar{p}_2) = \sigma^2(\bar{p}_1) + \sigma^2(\bar{p}_1)$ if sample is independent

$$\sigma(\bar{p}) = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$\sigma^2(\bar{p}_1 - \bar{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

$$\therefore \textit{the std error of } \bar{p}_1 - \bar{p}_2 \textit{ is } \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

This gives the test statistic:

$$\frac{(\overline{p_1} - \overline{p_2}) - 0}{\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}} \approx Z$$

Example

Party support went from 41% of 1500 to 45% of 1600. Is the increase in Recidivist party support statistically significant?

$H_0: p_2 \leq p_1$
$H_A: p_2 > p_1$

Let's us the p-value method…

$\overline{p_1} = 0.41 \; \overline{p_2} = 0.45$

$$\frac{(0.45-0.41)}{\sqrt{\frac{0.45(0.55)}{1600} + \frac{0.41(0.59)}{1500}}} = 2.25$$

p-value = P(Z>-2.25) = 1-0.9878=0.0122, which is <0.05

Reject H0. The increase in support is statistically significant.

Lecture 14/02/03

February 3, 2014
2:40 PM

Announcements

- No more office hours on Monday
- Homework: sample standard deviation (S_x) for samples (divide by n-1) is not the same as population standard deviation (σ_x) for whole population (divide by N).
- Homework: Question 4 - in $\bar{x} \pm Z\frac{S_x}{\sqrt{n}}$ x-bar is the statistic and the other half is the margin of error and the S_x/n \sqrt term is the estimate of the standard error of the statistic

Recapitulation

$$\frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx Z$$

$$\frac{(\overline{p_1} - \overline{p_2})}{\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}} \approx Z$$

Satterthwaite's method

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

OR $\min(n_1 - 1, n_2 - 2)$

Pooled variances and proportions - Proportions first

Suppose $\bar{p}_1 = \frac{x_1}{n_1}, \bar{p}_2 = \frac{x_2}{n_2}$

The pooled proportion is $\bar{p}_{pooled} = \frac{x_1 + x_2}{n_1 + n_2}$

And the test statistic is:

$$\frac{(\overline{p_1} - \overline{p_2})}{\sqrt{\frac{\bar{p}_{pooled}(1-\bar{p}_{pooled})}{n_1} + \frac{\bar{p}_{pooled}(1-\bar{p}_{pooled})}{n_2}}}$$

Example

Last week's poll: 615/1500 (RP) = 0.41
This week's poll: 720/1600 = 0.45

$H_0: p_2 \leq p_1$
$H_a: p_2 > p_1$

$H_0: p_2 \leq p_1$
$H_a: p_2 > p_1$

Method 1: no pooled proportion

$$Test\ statistic = \frac{0.45 - 0.41}{\sqrt{\frac{(0.45)(0.55)}{1600} + \frac{(0.41)(0.59)}{1500}}} = 2.250$$

We reject the null hypothesis. There is a significant increase

Method 2: p-bar pooled

$$\bar{p}_{pooled} = \frac{615 + 720}{1500 + 1600} = 0.430645$$

$$Test\ statistic = \frac{0.45 - 0.41}{\sqrt{\frac{(0.430645)(0.569355)}{1600} + \frac{(0.430645)(0.569355)}{1500}}} = 2.24977$$

We still reject the null hypothesis. There is still a significant increase.

Both are approximations of the Z-score even though p-bar pooled may be more accurate, but no matter what both are acceptable in ECON 227 and they are very close anyway. On tests, pick the one you like.

Next: Pooled Variances

The variation is the standard deviation squared:

$$Recall\ that\ S_1^2 = \left(\sqrt{\frac{\sum(x_{1_i} - \bar{x}_1)}{n_1 - 1}}\right)^2$$

… some algebra later …

$$S_{pooled}^2 = \frac{\sum(x_{1i} - \bar{x}_1)^2 + \sum(x_{2i} - \bar{x}_2)^2}{n_1 - 1 + n_2 - 1}$$

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_{pooled}^2}{n_1} + \frac{S_{pooled}^2}{n_2}}}$$

$$DF = n_1 + n_2 - 2$$

Example

	Women	Men
N	18	11
x-bar	121.3	119.2
s_x	12.3	15.8

Method 1: no pooled variance

$$\frac{121.3 - 119.2}{\sqrt{\frac{12.3^2}{18} + \frac{15.8^2}{11}}} = 0.376$$

Method 2: pooled variance

$$S_{pooled}^2 = \frac{17(12.3)^2 + 10(15.8)^2}{18 - 1 + 11 - 1} = 187.716$$

$$\frac{121.3 - 119.2}{\sqrt{\frac{187.716}{18} + \frac{187.716}{11}}} \cong 0.400$$

We don't have to do the pooled variance unless specifically asked. The keyword is ANOVA.

Intro to the Chi-squared Thing

$\chi^2 - Chi\ squared$

A chi-square random variable is of the form $Z_1^2 + Z_2^2 + Z_3^2 + \cdots + Z_i^2$

It is the sum of a finite number of independent Z^2 random variables.

Enter Pearson: Noodling with algebra - stats formulas mostly arose in the last 120 years while statisticians were playing around with formulas.

Poisson

$E(x) = \lambda, \sigma(x) = \sqrt{\lambda}$

$$\text{z-score} = \frac{x - \lambda}{\sqrt{\lambda}}$$

In the current notation X is called O (observed data)
Lambda is called E (expected frequency)

$$(z - score)^2 = \left(\frac{O - E}{\sqrt{E}}\right)^2 \pm \frac{(O - E)^2}{E}$$

Lecture 14/02/05

February 5, 2014
2:33 PM

Announcements

Chi-square might be on quiz - more information to follow

Chi-square

Pearson:
$$\frac{x-\lambda}{\sqrt{\lambda}} = \frac{number - mean}{stdev.}$$

(from Poisson)

Z-score, Poisson, all around the same time, Pearson did this:

$$\left(\frac{x-\lambda}{\sqrt{\lambda}}\right)^2 = \frac{(x-\lambda)^2}{\lambda}$$

We often use O (observed datum) for x and E (expected value) for the mean.

We get:
$$\frac{(O-E)^2}{E}$$

Pearson's test statistic:
$$\sum \frac{(O-E)^2}{E} = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

If the sample is big enough the sum of the observed minus the expected squared minus the expected does have approximately the same distribution as the sum of Z-squared. A random variable with the same distribution as a sum of independent Z-square random variables is said to have a chi-squared distribution. (on the back of the t-tables)

How big is big enough? It's complicated. Don't worry about it.

Example 1: The Professor X example

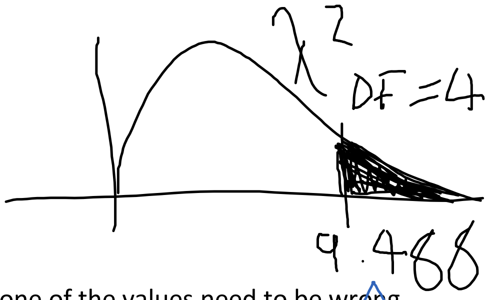
Professor X claims that 35% of his students get As, 40% get B, 15% get C, 5% get D and 5% get F. A skeptical student society has collected a SRS of 200 grades from Prof X. Here is the data:

	O	E
A	63	70
B	72	80
C	36	30
D	14	10
F	15	10
	200	200

Aside: If Professor X is making up the numbers and the size of the sample is fixed (say for 200), he can only make up numbers for the first 4 grades and the last grade is just 200-(numbers already chosen). Prof X is "free" to make up the first four grades, which is why it's called the degrees of freedom (there are 4 DF in this example).

Test whether the grade distribution is significantly different from that claimed by Prof X.

H_0: *the grade dist. is as Prof X claims*
H_A: *the grade dist. is significantly different from his claims*



OR

$H_0: p_A = 0.35; p_B = 0.40; p_C = 0.15; p_D = 0.05; p_F = 0.05$
H_A: *at least one of the proportions is different from Prof X's value*

NOT: $H_A: p_A \neq 0.35; p_B \neq 0.40; p_C \neq 0.15; p_D \neq 0.05; p_F \neq 0.05$ because only one of the values need to be wrong, not necessarily all.

Test statistic

$$= \sum \frac{(O-E)^2}{E} = \frac{(63-70)^2}{70} + \cdots + \frac{(15-10)^2}{10} = 6.8$$

We do not reject the null hypothesis. The grade distribution is not significantly different from the claim of Prof X.

<u>This is called a goodness-of-fit test.</u>

Example 2: Summer Drink preferences

A kiosk has been set up in a shopping mall with 3 popular summer drinks to be sampled. The kiosk believes that there is no significant difference in summer drink preference.

	Designer H2O	OJ	Cola
Preferred	83	126	141

Test whether there is a significant difference in drink preferences.

H_0: *there is no significant difference*
H_A: *there is a significant difference*

OR

$$H_0: p_{H20} = \frac{1}{3}; p_{OJ} = \frac{1}{3}; p_{cola} = \frac{1}{3}$$

$$H_A: \textit{at least one } p_k \textit{ is significantly different from } \frac{1}{3}$$

	Designer H2O	OJ	Cola	Total
O	83	126	141	350
E	116.666	116.666	116.666	350

Test statistic:

$$= \sum \frac{(O-E)^2}{E} = \frac{(83-116.666)^2}{116.666} + \frac{(126-116.666)^2}{116.666} + \frac{(141-116.666)^2}{116.666} = 15.537$$

We reject the null hypothesis. There is a significant difference in drink preference (at least one proportion is significantly different from 1/3.

Shortcut formula

$$\sum \frac{(O-E)^2}{E} = \left(\sum \frac{O^2}{E}\right) - n$$

Chi-square independence test

Two events are independent if: $P(A \cap B) = P(A)P(B)$

Example 3

The question is: Does the event that a randomly selected adult in Eurelia K smokes depend significantly on whether the person is M or F.

	F	M
Smokes	57	44
No smokes	243	156
	300	200

We could just see if $\overline{p_1} - \overline{p_2}$ is not 0.

Approach 1

$H_0 = p_F = p_M$
$H_a = p_F \neq p_M$

Test statistic

$$= \frac{0.19 - 0.22}{\sqrt{\frac{(0.19)(0.81)}{300} + \frac{(0.22)(0.78)}{200}}} = -0.81$$

We don't reject the null hypothesis. There is no significant difference in the proportion of smokers between F and M in Eurelia K.

Lecture 14/02/10

February 10, 2014
2:33 PM

Back to the smokers

	F	M	
Smokers	57	44	101
Non-smokers	245	156	399
	300	200	500

contingency table

Approach 1

$H_0 = p_F = p_M$
$H_a = p_F \neq p_M$

Test statistic

$$= \frac{0.19 - 0.22}{\sqrt{\frac{(0.19)(0.81)}{300} + \frac{(0.22)(0.78)}{200}}} = -0.810219194$$

We don't reject the null hypothesis. There is no significant difference in the proportion of smokers between F and M in Great Eurelia. (specify where this sample came from... Eurelia is not the world.

Approach 2: Pooled proportions (inexplicably preferred by most textbooks)

$H_0 = p_F = p_M$
$H_a = p_F \neq p_M$

Before doing anything else, calculate the so called pooled proportion.

$$\bar{p}_{pooled} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{101}{500} = 0.202$$

Test statistic

$$= \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\frac{\bar{p}_{pooled}(1-\bar{p}_{pooled})}{n_1} + \frac{\bar{p}_{pooled}(1-\bar{p}_{pooled})}{n_1}}}$$

$$= \frac{0.19 - 0.22}{\sqrt{\frac{(0.202)(0.798)}{300} + \frac{(0.202)(0.798)}{200}}} = -0.818530275$$

Why do we pool proportions? Because of approach 3.

Approach 3: Pearson and Chi-squared

Recall statistical independence formula:

$$P(A \cap B) = P(A)P(B)$$

Statisticians used to play with formulas like this a lot. What Pearson did was, instead of asking whether there is a significant difference between proportions of male and female smokers, we ask whether there is statistical independence between the event being a smoker and the event of being a female or male.

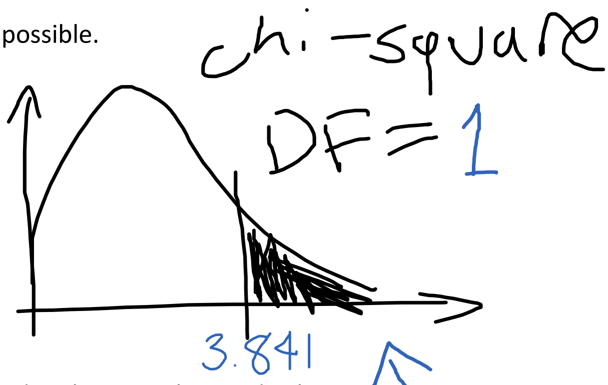
H_0: *the event of being a smoker is independent of whether the person is female or male*
H_A: *the event of being a smoker depends on whether the person is female or male*

H_0: *the event of being a smoker is independent of whether the person is female or male*
H_A: *the event of being a smoker depends on whether the person is female or male*

These are always right-tailed tests for us, although left-tailed tests are possible.

	E_11	E_12	101
	E_21	E_22	399
	300	200	500

Contingency table



Contingency tables come from insurance companies.

Pearson calls the cells of the table expected frequencies to express the idea they are observed values.

An aside from Kmack: don't believe everything you read on the Google.

Pearson's idea: choose E_ij in such a way that the independence criterion holds in each cell.

Kmack has randomly chosen the top right cell - E_12

$$P(S \cap M) = P(S)P(M)$$
$$\frac{E_{12}}{500} = \left(\frac{101}{500}\right)\left(\frac{200}{500}\right)$$
$$E_{12} = \frac{(101)(200)}{500} = 40.4$$

$$E_{ij} = \frac{(row\ sum)(column\ sum)}{n} = \frac{R_i C_j}{n} = \frac{RC}{n}$$

This formula for the expected frequency is true no matter which cell you choose.

	60.6	**50.4**	101
	239.4	*149.6*	399
	300	200	500

After filling in the first cell, there is no more freedom for what numbers go in the other cells because the totals are already known (i.e. 101-50.4=60.6). There is only one degree of freedom, DF=1.

Test statistic
$$= \sum \frac{(O-E)^2}{E}$$
$$= \frac{(57-60.6)^2}{60.6} + \cdots + \frac{(156-149.6)^2}{149.6} = 0.669991811$$

Again, we do not reject the null hypothesis.

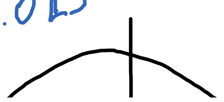
Conclusion: The event of being a smoker is independent of whether the person is female or male or, more accurately, smoking is not significantly dependent on gender.

Test statistic from approach 2: $(-0.818530275)^2 = 0.669991811$

This is the reason why textbooks want you to use pooled proportions because the square of the test statistic from the pooled approach is the test statistic from the chi-squared approach.

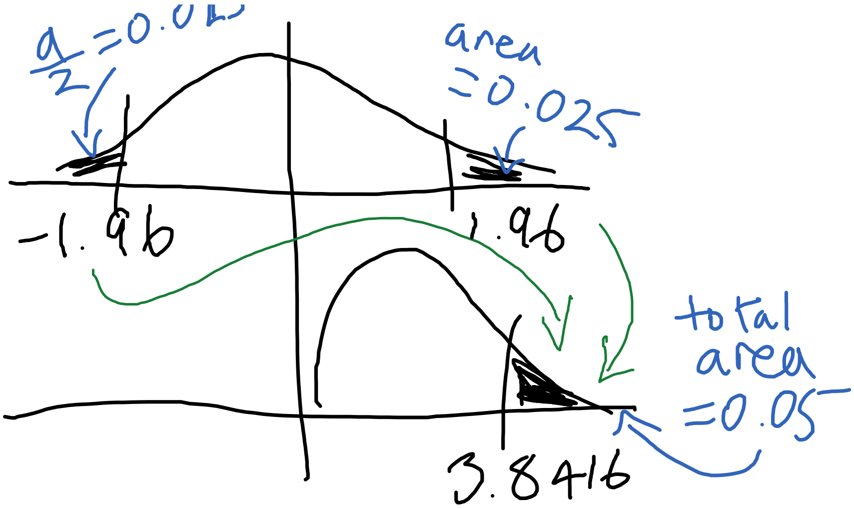
$H_0: p_1 = p_2$
$H_A: p_1 \neq p_2$

$\frac{\alpha}{2} = 0.025$



area =0.

$H_0: p_1 = p_2$
$H_A: p_1 \neq p_2$



Larger Contingency-Table Example

Income\Age	A_1 <30	A_2 30-49	A_3 50-64	A_4 >=65	
I_1 < 35K	25	11	6	10	52
I_2 35K - 75K	12	32	4	11	59
I_3 >75K	3	10	11	15	39
	40	53	21	36	150

Independent always goes in the null dependent always goes in the alternative

H_0: *Age and income are independent*
H_A: *Age and income are dependent*

	$\frac{52 \times 40}{150} = 13.86$				

DF = 6

Quiz on Feb 19th for this section, Feb 17th for morning section

Lecture 14/02/12

February 12, 2014
2:39 PM

Announcements

- Quiz is on the 24th of February for the afternoon section, Kmack was confused.
- Quiz is 50 mins long, material is the same as the 20 questions material, z-test, p-test.

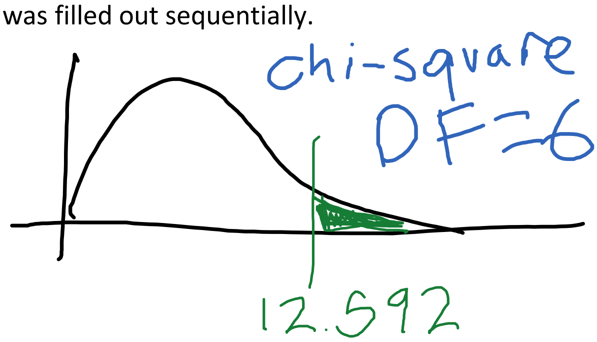
Back to the Larger Contingency-Table Example

Income\Age	A_1 <30	A_2 30-49	A_3 50-64	A_4 >=65	
I_1 < 35K	25	11	6	10	52
I_2 35K - 75K	12	32	4	11	59
I_3 >75K	3	10	11	15	39
	40	53	21	36	150

Degree of freedom really means something here because as you fill out the table you will have no choice but to fill in a number because all cells must add up to the cell and row totals.

	13.9	18.4	7.3	*12.5*	52
	15.7	20.8	8.3	*14.2*	59
	10.4	*13.8*	*5.5*	*9.4*	39
	40	53	21	36	150

Italicized numbers had no freedom in choice when the table was filled out sequentially.
m
$H_0: Age\ and\ income\ are\ independent$
$H_A: Age\ and\ income\ are\ dependent$



This is a chi-squared test.

The degree of freedom is always:
$DF = (rows - 1)(columns - 1) = (3 - 1)(4 - 1) = 6$

Test statistic:

$$= \sum \frac{(O - E)^2}{E}$$

$$= \frac{(25 - 13.9)^2}{13.9} + \cdots + \frac{(12 - 15.7)^2}{15.7} + STOP!!$$

Tip from Kmack: keep a running total because sometimes you might exceed your critical chi-square value before summing up all the cells. In this example, the moment you get into the rejection region by exceeding 12.592 you can stop computing the sum.

Tip from someone else: You can go even faster if you look at the contingency table and pick out the biggest values. In our example we could have crossed the critical value with only two cells.

Since the test statistic is already > critical chi-square (i.e. 12.592) we reject the null hypothesis. Age and income are significantly dependant.

Shortcut from Kmack:

$$test\ statistic = \sum \frac{(O-E)^2}{E} = \left(\sum \frac{O^2}{E}\right) - n$$

In this example:

$$= \frac{25^2}{13.9} + \cdots + \frac{15^2}{9.4} - 150 = 69.37$$

So we would have been well into the rejection region.

Summary so far

1. Prof X type (**G**oodness **O**f **F**it test)
2. Summer drink example (GOF test) - we don't say what the probability distribution is, but the knowledge of what our goal is allows us to infer if there is no significant difference between how the 3 drinks are like then the probability of being liked should be around 1/3 for each.
3. Chi-squared (contingency table)

GOF Tests

1. The null hypothesis always says that the data have been sampled from a population with some stipulated probability distribution. The alternative hypothesis says the data comes from a population with some other distribution.
2. Next, the data are categorized by a partition, namely K events that are mutually exclusive and collectively exhaustive.
3. Then the observed frequency (O) are counted, giving the number of data in each category.
4. Ensuite, the expected frequencies are $E_j = np_j, where\ p_j$ is the probability for each category, according to the probability distribution in H_0
5. $DF = K - m - 1$ (in ECON 227, m will always be 0).
   K is the number of categories (or "Kategories").
   M is the number of parameters that have to be estimated using the sample data. It will always be 0 in ECON227.

Forensic Chi-square

In ECON 227, every chi-squared test on any question we do will be a right tailed test.

Can there be left tailed tests? Yes.

Mendel (Brno, Czech republic) was an abbot who lived around the same time as Darwin. He cross-pollenated a bunch of flowers and figured out you could predict the proportions of traits like flower colour. People who knew the science at the time were suspicious that his data was too accurate.

Kmack asks: has anyone heard of the book *The Case of the Midwife Toad*? Essentially the data was too good.

$$\sum \frac{(O-E)^2}{E}$$

If your observed is too close to the expected, your test statistic will be close to 0. Using a left tailed test, statisticians can "forensically" investigate whether data is too good and recommend other scientists try to replicate the experiment.

But again, no question on ANY test in this course will involve a left-tailed or two-tailed chi-square.

Lecture 14/02/17

February 17, 2014
2:48 PM

Kmack tries to use the overhead

- Kmack will post the ANOVA page on myCourses or make enough copies for everyone next class.
- Kmack will go through the quiz the morning section wrote on Wednesday
- Kmack has put up solutions to Quiz 3 on myCourses.

Analysis of variance or… ANOVA

Kmack tells a story.

You are driving in the car with your friend. Kmack has a bad habit where he doesn't check the fuel level, but his friend in the passenger seat eyes the needle approaching red with concern. Kmack says he will fill up, but drives past a White Rose gas station because he dislikes the sludge in White Rose gas. His passenger laughs at Kmack for avoiding White Rose gas because they use the same gas as everyone else, in fact he has even seen a tanker truck fill up a White Rose station and then a shell station.

The passenger is saying all of them are equal, but the driver is saying there's at least on that's different.

Let's suppose mu represents average litres of gas per 100 km of driving range.

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
$H_A: at\ least\ one\ \mu_j, is\ significantly\ different$

The theoretical starting point for these kinds of computations.

1. The samples are taken from normally-distributed populations.
2. The samples are randomly and independently selected.
3. All the variances of the populations are the same.
   $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 = \sigma^2$

The test is based on two different estimates of the homoscedastic σ^2

In statistics, a sequence or a vector of random variables is homoscedastic if all random variables in the sequence or vector have the same finite variance. This is also known as homogeneity of variance.

The first one is simply a (weighted) average of all the sample variances. If all the sample sizes are the same, it is simply the ordinary mean of the S_x^2 values

	Shill	Canbas	Crindol	Petrol	Pumpex
Litres per 100km	9.6 9.1 9.5 9.9 9.1 9.7	10.2 …	16.3 …	9.8 …	9.2 …
$\bar{x}_j$	9.4833	9.7	15.766	9.5	9.4
S_j^2	0.10566	0.164	0.090667	0.14	0.172

Kmack cooked up the data so that we will reject the null hypothesis, but notice that the standard deviations-squared are not too different from each other.

First estimate of σ^2

$$= \frac{S_1^2 + \cdots + S_5^2}{5}$$

$$= 0.134466533$$

This is called MSE (error mean square) and tends to be a good estimate of σ^2 regardless of whether the null hypothesis is true or false.

The second estimate of σ^2 is based on the formula $\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n_j}}$ ie.

$$\sigma^2(\bar{x}) = \frac{\sigma^2}{n_j}$$

$$\sigma^2 = n_j \sigma^2(\bar{x})$$

We estimate σ^2 by

$$n_j \times stdev\ of\ \{\overline{x_1}, \ldots, \overline{x_k}\}$$

This estimate is called the treatment mean square (MSTR). MSTR tends to be smaller than the first estimate if the null hypothesis is true and larger than the first estimate if the null hypothesis is false.

In our gasoline example,

$$MSTR = 6 \times (Sample\ std\ dev\ of\ \{9.48333, 9.7, 15.76666, 9.5, 9.4\})^2$$

$$MSTR = 6 \times (7.814222222) = 46.885333333$$

MSE tends to be good estimate whether the null hypothesis is true or false. Here the null hypothesis is false because 46.885333333 is larger than our first estimate of 0.134466533

The test statistic is always $\frac{MSTR}{MSE}$

$$= \frac{46.885333333}{0.134466533} = 348.7$$

This is a huge test statistic.

The appropriate tables to be used with this statistic are the F-tables which we have not yet received, but will receive Wednesday.

We shall see on Wednesday that the test statistic is well into the rejection region. We reject the null hypothesis. Thus, at least one of the means is significantly different.

This was a lot of computational work. There is a shortcut - the computations will be easier when we use the ANOVA tables

Side comment

$$H_0: \mu_1 < \mu_3$$

$$H_A: \mu_1 \neq \mu_3$$

Apply the ANOVA technique to columns 1 & 3

	Shill	Crindol
Litres per 100km	9.6 9.1 9.5 9.9 9.1 9.7	16.3 …
$\bar{x}_j$	9.4833	15.766
S_j^2	0.10566	0.090667

$$MSE = \frac{S_1^2 + S_3^2}{2} = 0.0981666666$$

$$MSTR = 6 \times (sample\ std\ dev\ of \{9.48333333, 15.7666666\})^2$$
$$= 6 \times 19.74013885 = 118.44083$$

$$\frac{MSTR}{MSE} = 1206.528019$$

Let's compare this with the pooled variance method

$$S_{pooled}^2 = \frac{(n_1 - 1)S_1^2 + (n_3 - 1)S_3^2}{n_1 + n_3 - 2} = 0.0981666666 = MSE$$

The pooled variance-squared is just the MSE from the ANOVA method

Test statistic

$$= \frac{\bar{x}_1 - \bar{x}_3}{\sqrt{\frac{S_{poooled}^2}{n_1} + \frac{S_{pooled}^2}{n_3}}} = -34.735427$$

$$(-34.753427)^2 = 1206.52802 = \frac{MSTR}{MSE}$$

The pooled variance test statistic squared is just the ANOVA test statistic.

Lecture 14/02/19

February 19, 2014
2:34 PM

Announcements

- Kmack passes out prints of analysis of variance sheets and F-tables (variance sheet on MyCourses)

One-Way ANOVA

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$
H_A: $at\ least\ one\ \mu_j\ is\ different$

Two estimates of σ^2: MSE & MSTR

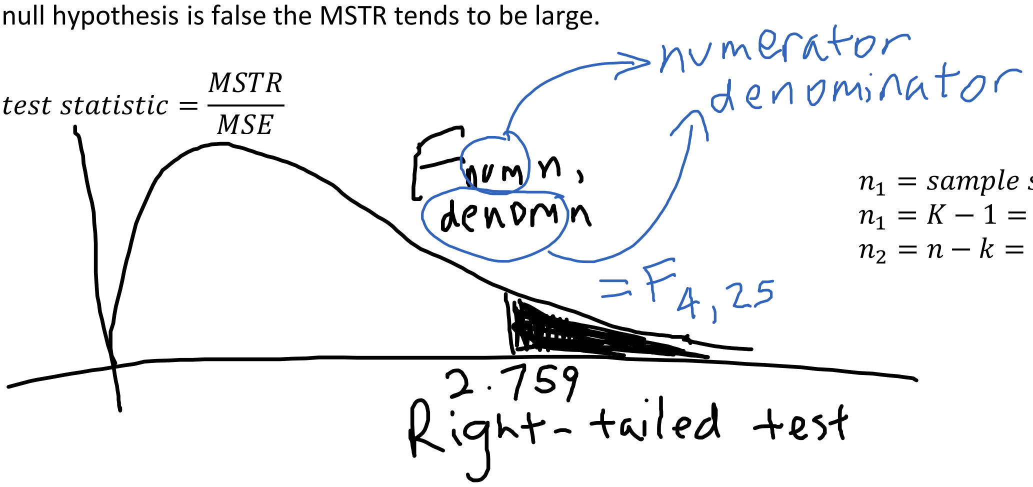
$MSE = ordinary\ or\ weighted\ average\ of\ s_1^2, {s_2}^2, s_3^2, \ldots, s_k^2$

The weighted average is used if the sample sizes are not the same.

$MSTR = n_j \times estimated\ SE\ of\ the\ \bar{x}\ values$

If the null hypothesis is true, MSE tends to be a good estimate of σ^2 whether or not the null hypothesis is true. If the null hypothesis is false the MSTR tends to be large.

$test\ statistic = \dfrac{MSTR}{MSE}$



$n_1 = sample\ size\ for\ MSTR$
$n_1 = K - 1 = num\ DF$
$n_2 = n - k = denom\ DF$

We use the 0.05 for the one-way ANOVA, two-way ANOVA, and regression testing.

For the gasoline example:

Test statistic
$= \dfrac{46.885333}{0.1344667}$
$\approx 349 \gg 2.759$

Reject the null hypothesis. At least one type of gas has a significantly different mean L/100km

Review of the morning section's Feb 17 quiz

1. a) 90% CI for the mean amount on the debit cards

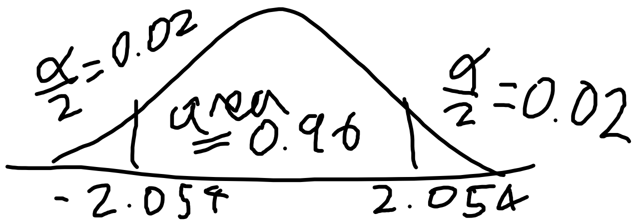
   amount: 5, 25, 10, 5, 15, 50, 10, 15, 5, 5

$\bar{x} = 14.5$
$s_x = 14.0337$
$n = 10$

$t_{DF=9}$: 1.833 $(0.05\ column)$

$$145 \pm \frac{1.833(14.0337)}{\sqrt{10}} = 14.5 \pm 8.1$$

b)
Income: 20, 35, 30, 30, 25, 60, 30, 35, 25, 20

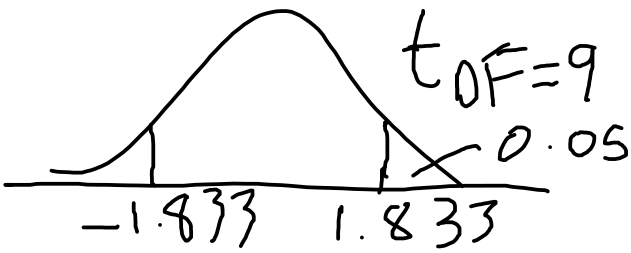


$S_x = 11.4988$

$$n = \frac{(2.054)^2(11.4988)^2}{(0.5)^2} = 2232$$

$-2.05: 0.0202$
$-2.06: 0.0197$

c) Test at the 10% level of significance



$H_0: \mu = 10$
$H_A: \mu \neq 10$

Test statistic
$$= \frac{14.5 - 10}{\frac{14.0337}{\sqrt{10}}} = 1.014$$

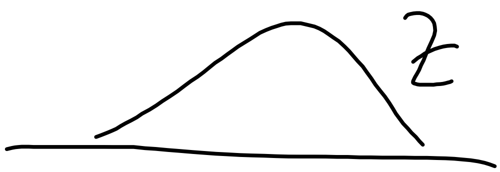
We do not reject the null hypothesis. The mean amount is not significantly different from $10.

d) There aren't enough numbers in the table to get close enough. Sophisticated answer: the numbers on the t-table are too sparse to get a very good approximation.

2.

	MTL	$\overline{MTL}$	Total
F	488	112	600
M	312	88	400
Total	800	200	1000

a) Test whether the proportion of female students significantly exceeds 57%



$H_0: p \leq 0.57$
$H_A: p > 0.57$

$\bar{p} = 0.6$

Nothing you calculate from the data should ever go into the null hypothesis.

Test statistic

$$= \frac{0.6 - 0.57}{\sqrt{\frac{(0.57)(0.43)}{1000}}} = 1.92$$

Aside: We should use:

$$\frac{\bar{p} - p}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}}$$

$$p - value = P(Z > 1.92) = 1 - 0.9726 = 0.0274$$

Reject the null hypothesis because p-value < 0.05

b) Estimate the standard error of the sample proportion of female students in the group from MTL

$$\frac{488}{800} = 0.61$$

$$\sqrt{\frac{(0.61)(0.39)}{800}} \approx 0.017$$

c) We don't know the proportions here, use 0.5 and 0.5 because we don't know anything about dissatisfied students

$$n = \frac{Z^2 p(1-p)}{E^2}$$
$$= \frac{2.576^2 (0.5)^2}{(0.008)^2} = 25921$$

d)

$$0.8 \pm 2.576\sqrt{\frac{(0.8)(0.2)}{1000}} = 0.8 \pm 0.0325$$

e) Conservative formula is the worst case scenario - it is the most you will need to ask for that level of confidence.

Lecture 14/02/26

February 26, 2014
2:38 PM

Announcements

- Next quiz sessions: Thursday 2:30 LEA517 and Friday 11:30 LEA424

Recap of Gasoline Example

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
$H_A: at\ least\ one\ \mu_j\ is\ different$

$$MSE = \frac{S_1^2 + S_2^2 + S_3^2 + S_4^2 + S_5^2}{5}$$

Kmack wants us to understand how ANOVA works by explaining MSE, but memorizing the steps also works.

The MSE tends to be a good estimate of σ^2 whether the null hypothesis is true or false. MSTR tends to be too big if the null hypothesis is false (recap, we already looked at this).

$$\frac{MSTR}{MSE}$$

Has F with $numDF = K - 1$ and $denomDF = n - K$

$n = n_1 + n_2 + \cdots + n_k = overall\ number\ of\ data$

$K = number\ of\ different\ samples\ (treatments)$

This was first applied to agriculture and fertilizer, hence the term "treatment".

For the gasoline example, MSE = 0.1344667, MSTR = 46.885333, test stat = 349

If the sample sizes are not all the same, we must use the weighted average.

$$\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \cdots + (n_k - 1)S_k^2}{n_1 - 1 + n_2 - 1 + \cdots + (n_k - 1)}$$
$$= \frac{\sum (n_j - 1)S_j^2}{n - k}$$

In the ANOVA context, we will always estimate σ by $\sqrt{MSE}$

$CI\ for\ \mu_j$
$$\bar{x} \pm t\frac{\sqrt{MSE}}{\sqrt{n_j}}$$

For $\mu_i - \mu_j$
$$\bar{x}_i - \bar{x}_j \pm t\sqrt{\frac{MSE}{n_i} + \frac{MSE}{n_j}}$$

If Kmack asks for an ANOVA confidence interval you will only get full marks if you use the sqrt-

MSE formula

We shall now do ANOVA the way the rest of the world does it. This is called an ANOVA table.

ANOVA Table: Gasoline Example
(Sum of squares)

Source	DF	SS	MS	F
Treatment	k-1	SSTR	MSTR	$\frac{MSTR}{MSE}$
Error	n-k	SSE	MSE	
Total	n-1	SSTOTAL		

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

	9.6 9.1 9.9 9.1 9.7 9.5	10.2 9.9 9.5 9.1 10.0 9.5	16.3 15.4 15.7 15.8 15.6 15.8	9.8 9.5 9.9 9.0 9.1 9.7	9.2 9.0 8.9 9.8 9.8 9.7
T_j	56.6	58.2	94.6	57	56.4

$$SSTOTAL = 3670.69 - \frac{323.1^2}{30} = 190.903$$

Source	DF	SS	MS	F
Treatment	4	187.5413	46.88533	~349
Error	25	3.361667	0.134466	
Total	29	190.903		

$$SSTR = \left(\sum \frac{T_j^2}{n_j}\right) - \frac{(\sum x)^2}{n}$$
$$= \frac{56.9^2}{6} + \frac{58.2^2}{6} + \frac{94.6^2}{6} + \frac{57^2}{6} + \frac{56.4^2}{6} - \frac{323.1^2}{30}$$

$$MS = \frac{SS}{DF}$$

The ANOVA table is from when people didn't have computers. Now Kmack will demonstrate it on excel.

On Excel you will need to type out all the data the use the data analysis tool. This tool is an add-in that you will need download for free: google "data analysis toolpak". You will then need to run ANOVA single factor in the toolpak and select the data you entered. This tool will generate the ANOVA table from your data automatically.

Kmack will put up problems on mycourses if you want to practice during the reading week.

Lecture 14/03/10

March 10, 2014
2:42 PM

Announcements

- All quizzes are marked, Kmack is preparing the excel file for MyCourses. You may see him in his office hours after today to see the quiz.
- Tomorrow's office hours are moved to begin at 2pm.

ANOVA

$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$

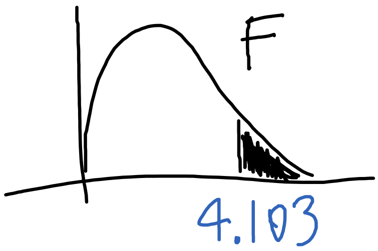
The way on the sheet on MyCourses is the long way of doing ANOVA

What is two way ANOVA?

The ANOVA we have seen is called one-way ANOVA, or some books like to call this method completely randomized design. Kmack likes one-way ANOVA and calling it "treatment" in reference to treating farms with fertilizer.

Warm-up example: one-way ANOVA

	A	B	C
	12	16	16
	23	31	23
	18	14	45
	25	8	
		12	
		9	
Totals	78	90	84



$H_0: \mu_A = \mu_B = \mu_C$
$H_A: at\ least\ one\ \mu_j\ is\ significantly\ different$

The ANOVA table

Source	DF	SS	MS	F-test statistic
Treatments	k-1=2	338.077	169.038	169.038/91.1=1.8555
Error	n-k=10	911	91.1	
Total	n-1=12	1249.077		

$$SSTOTAL = \left(\sum x^2\right) - \frac{(\sum x)^2}{n} = 6134 - \frac{252^2}{13} = 1249.077$$

$$SSTR = \left(\sum \cdot \frac{T_j^2}{n_j}\right) - \frac{(\sum x)^2}{n} = 338.077$$

The F-table has two sides, 0.05 and 0.025. This is a not a two tailed test so we use the 0.05 side.

CI formulas in the ANOVA context

$$x_i \pm t\frac{\sqrt{MSE}}{\sqrt{n_j}}$$

$$DF = error\ DF$$

$$x_1 - x_2 \pm t\sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

Part "b)": Give a 95% CI for the mean of μ_A population A.

$$\frac{78}{4} \pm \frac{2.228\sqrt{91.1}}{\sqrt{4}}$$ $t_{DF=10}$

$$19.5 \pm 10.6327$$

Where do the DF numbers come from?

	DF
k-1	Treatments - 1
n-k	Data - treatments
n-1	Data - 1

Kmack can't fit in a smart car.

Two-way ANOVA

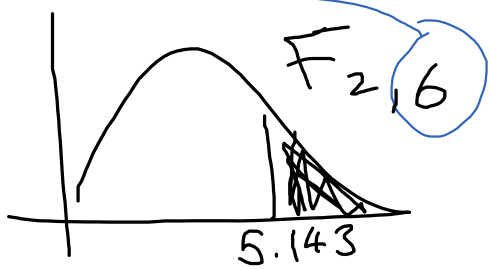
This next explains when a statistician says "we need to remove the effect of the car size from their fuel consumption". This is also known as "randomized block design".

Blocks\Treatments	Gas A	Gas B	Gas C	Total_i
Compact	9.1	8.9	7.8	25.8
Medium	10.2	11.1	9.3	30.6
Full	10.0	11.6	10.9	32.5
GUV	12.3	12.0	11.1	35.4
Total_j	41.6	43.6	39.1	124.3

$$H_0: \mu_A = \mu_B = \mu_C \ (after\ the\ effect\ of\ car\ size\ has\ been\ accounted\ for)$$
$$H_A: at\ least\ one\ \mu_j\ is\ different\ after\ the\ effect\ of\ car\ size\ has\ been\ removed$$

We usually don't write the extra words above but they will come up during the conclusion where you will state whether or not you rejected the null hypothesis after accounting for car size (in this case).

Source	DF	SS	MS	F-test statistic
Treatments	k-1=2	2.542	1.271	1.271/0.354=3.59
Blocks	Blocks-1=3	16.263	5.421	5.421/0.354=15.31
Error	treatmentsDF * blocks DF=6	2.125	0.354	
Total	n-1=11	20.929		



$$SSTOTAL = (\sum x^2) - \frac{(\sum x)^2}{n} = 1308.47 - \frac{124.3^2}{12} = 20.929$$

$$SSTR = \left(\sum \frac{T_j^2}{n_j}\right) - \frac{(\sum x)^2}{n} = 2.542$$

$$SSBLOCKS = \left(\sum \frac{T_i^2}{n_i}\right) - \frac{(\sum x)^2}{n} = 16.2325$$

The blocks f-statistic is big enough to be in the rejection region, but this means that there is a difference because of car sizes.

The treatments test statistic is 3.59 so we do not reject the null hypothesis. The mean L/100km are not significantly different for the 3 kinds of gas, after the effect of car size has been accounted for.

Lecture 14/03/12

March 12, 2014
2:40 PM

Announcements

- 10-12 tomorrow in Kmack's office if you want to pick up your quiz
- Excel grades will be up on mycourses by tomorrow

Today:

Using excel to calculate two-way ANOVA

Enter data table into excel:

Blocks\Treatments	Gas A	Gas B	Gas C
Compact	9.1	8.9	7.8
Medium	10.2	11.1	9.3
Full	10.0	11.6	10.9
GUV	12.3	12.0	11.1

We are doing Two-Factor ANOVA without replication (ie. There is only one data point in each cell). In the real world there will usually be multiple data points.

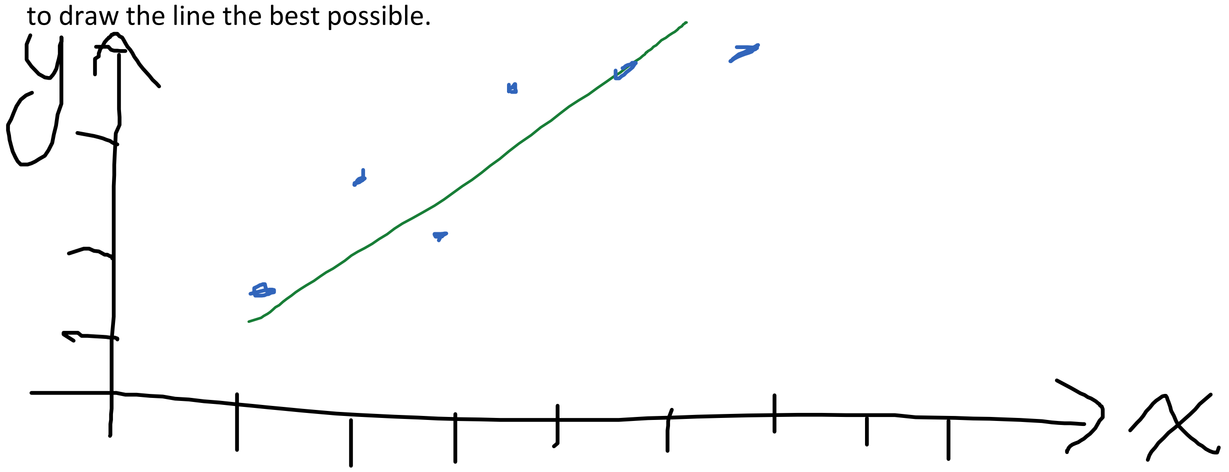
Highlight the table and run the analysis. You should get an ANOVA table where rows are car types and columns are gas types.

Also in this table there is the p-value with which we can reject the null hypothesis simply by comparing the p-value to our significance threshold (0.05)

Last major of the topic: Regressions

We begin with **simple regression.**

There is also multiple regression, linear regression, logistic regression and much more! The point is that you have a bunch of points on a graph and you want to draw a line the goes through those points. Usually you won't be able to draw a line through every point, but the objective is to draw the line the best possible.



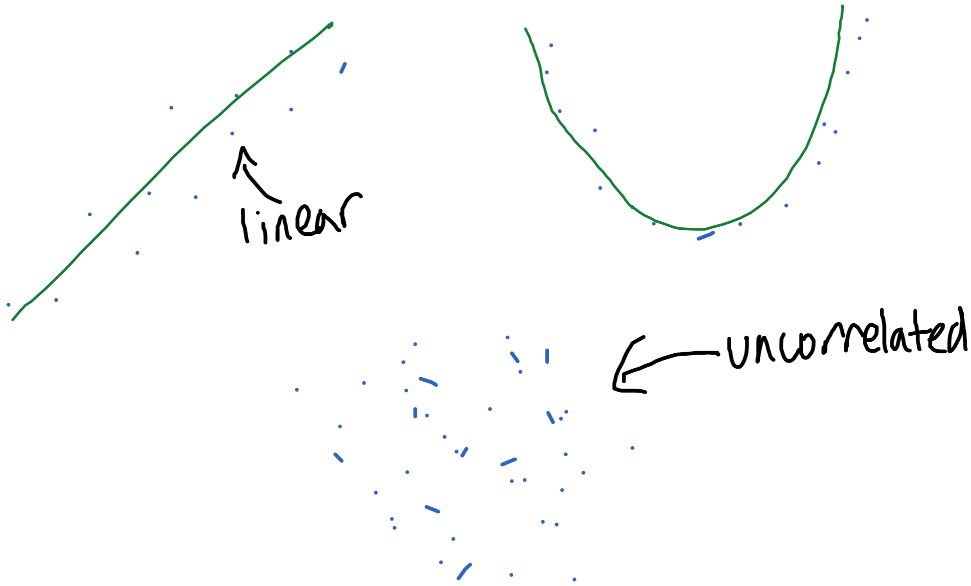
Simple Linear Regression

An example: we hope that the more you spend on advertising, the more you get back in revenue.

We construct a scatter plot where x = monthly advertising expenditure and y = corresponding monthly revenue.

If the points of our scatter plot are clustered around a straight line, it is often useful to estimate an equation for the line.

Points don't always have to cluster around a *straight* line. You could also have things like this:



Why do we call this regression?

Dalton (cousin of Darwin) did this stuff with data, one of which involved how often Prussian soldiers were kicked by their horses. His earliest work began with the idea of graphing data from things we think to be true, such as tall parents have children. Dalton found that children of tall parents tended to be tall, but shorter than their parents and the children of short parents tended to be short, but taller than their parents. Hi remark was that children were *regressing* towards the mean.

Here is a trick question

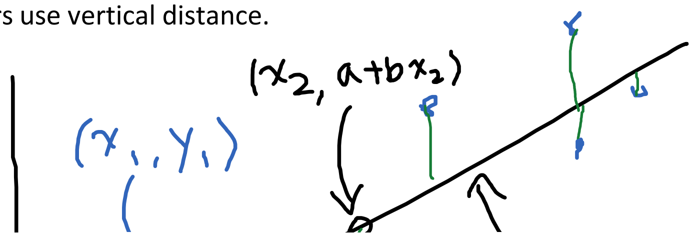
Would you get a good line by sticking a meter stick onto a scatter plot? Probably yes, apparently the human eye is good for this kind of thing.

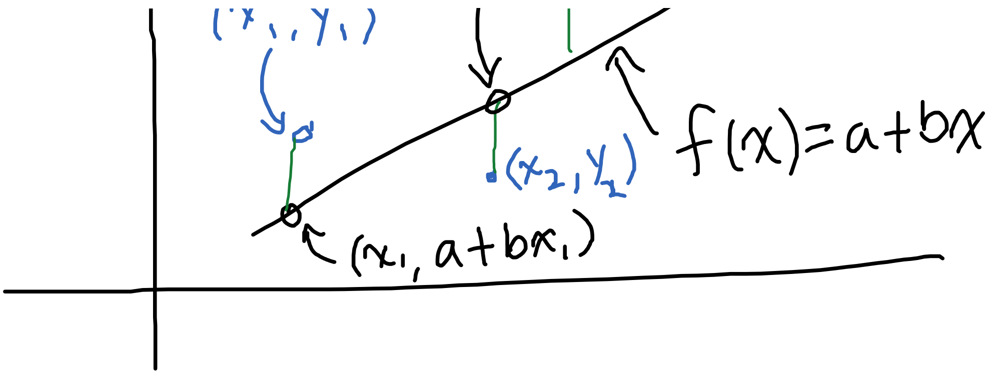
So why do we need a formula? Because it would be silly for statisticians to be wiggling meter sticks all day.

Today we talk about least-squares (line of best fit)

Let's suppose we have a scatter plot and a LOBF, and the equation of the line is $f(x) = a + bx$ ... most calculators use a+bx and most people grew up with mx+b. These are all equivalent.

Although we think of shortest distance between a point and the line as the perpendicular distance, most calculators use vertical distance.





$y_1 - (a + bx_1) = y - a - bx_1$

Is the vertical distance of the first point from the line. This expression is called the residual e_1

$y_2 + (a + bx_1) = y - a - bx_2$

Is the negative vertical distance from the second point to the line, this distance would work out to be negative because it is below the line, and it is called e_2 the second residual.

Since residuals can be positive or negative, steps are taken to prevent the negative residuals from cancelling out the positive residuals. It is possible to take the absolute possible, but in the case of least-mean-square, we square it.

The ordinary least squares line (OLS) is the line for which the average of the squares of the residuals is minimized. That is: $\frac{e_1^2+e_2^2+e_3^2+\cdots+e_n^2}{n}$ is as small possible.

The values of a and b that minimize this average are the same values that minimize the sum of the squares of the residuals.

The technique used here with calculus is to minimize by setting the derivative to zero, but we will not be tested on this in ECON 227.

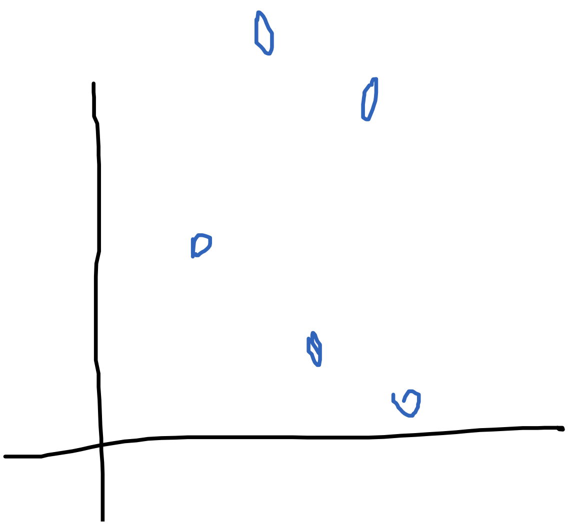
Here are some wild formulas derived using calculus

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$a = \frac{(\sum y) - b(\sum x)}{n}$$

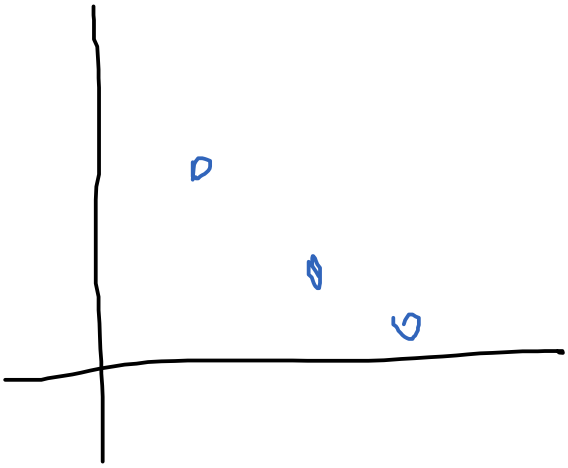
Example:

x_i	y_i
12	42
24	58
31	7
18	88
18	21



n=5 data points

x_i	y_i
12	42
24	58
31	7
18	88
18	21



n=5 data points

$\sum xy = 3075$
$\sum x^2 = 2329$
$\sum x = 103$
$\sum y = 216$

$$b = \frac{5(3075) - (103)(216)}{5(2329) - 103^2} = -6.634169884$$

$$a = \frac{(216) - (-6.634169884)(103)}{5} = 179.98638996$$

Your calculator can also do this, the data is usually entered the same way. However with this example something is wrong. Don't use this example except to see which numbers go where. $\sum xy = 3075$ should be $\sum xy = 4075$

Lecture 14/03/19

March 19, 2014
2:37 PM

Announcements

- Afternoon is now behind morning because of Kmack's eye-drops, still lots of time to review.
- Second assignment will likely come out this week.

This example again

	x_i	y_i	$\sum xy$	$\sum x^2$
	12	42	504	144
	24	58	1392	576
	31	7	217	961
	18	88	1584	324
	18	21	378	324
Total	103	216	4075	2329

$n = number\ of\ data\ points = 5$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

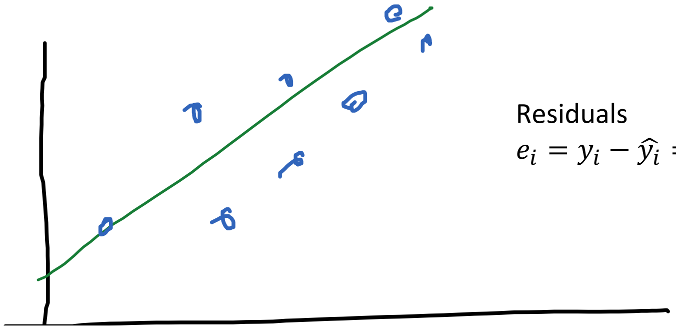
$$a = \frac{(\sum y) - b(\sum x)}{n}$$

$\sum xy = 4075$
$\sum x^2 = 2329$
$\sum x = 103$
$\sum y = 216$

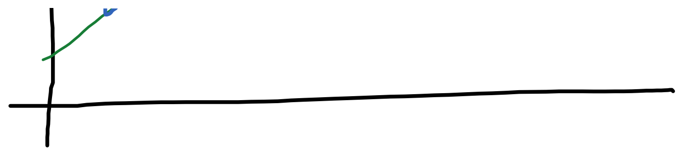
$$b = \frac{5(4075) - (103)(216)}{5(2329) - 103^2} = -1.807915058$$
$$a = \frac{(216) - (-1.807915058)(103)}{5} = 80.44305019$$

$\hat{y} = a + bx$
$\hat{y} = 80.443 - 1.808x$

Work this out using the calculator: input data like before for two-variable ([x] [comma] [y] [M+] for Sharp/Casio, two lists for TI-8X).



Residuals
$e_i = y_i - \hat{y}_i = y_i - (a + bx_i)$



	x_i	y_i	$\hat{y}_i$	e_i
	12	42	58.748	-16.748
	24	58	37.053	20.947
	31	7	24.398	17.398
	18	88	47.901	40.099
	18	21	47.901	-26.901
Total				-0.001

There is a bit of rounding error:
$\hat{y}_1 = \hat{y}(12) = 58.748$
$= 80.443 - 1.808(12) \approx 58.747$

The sum of the residuals should add up to essentially zero (within our rounding error).

The Folklore of Regression
As told by Great Bard MacKenzie

Say we have a bunch of dots on a scatter plot that looks like they all cluster around a straight line. Our hope is that the line passes as close as possible to the dots but the statistician will notice the dots appear above and below the line, in other words there is some **e**rror (e_i), or rather, residuals. The statistician will call the sum of the squares of the residuals the unexplained variation in the y-values.

$$\sum e_i^2 = sum\ of\ the\ squres\ of\ the\ residuals = unexplained\ varition\ in\ the\ y - values.$$

This unexplained variation is also called the **e**rror **s**um of **s**quares (SSE).

Total sum of squares
$SSTOTAL = \sum(y - \bar{y})^2$

Way back when we learned about sample standard deviation and population standard deviation. Recall that:

$$S_y = \sqrt{\frac{\sum(y - \bar{y})^2}{n-1}}$$

The two look alike and it's because both a measures of how spread out data is. In fact, the two are related in this way:

$$S_y^2 = \frac{\sum(y - \bar{y})^2}{n-1}$$
$$S_y^2(n-1) = \sum(y - \bar{y})^2$$

They wanted to call this variance but that term was already used by Dalton so they picked a new name: variation. As such, the folkloric term for SSTOTAL is: the total variation in Y.

either formula works

For our current dataset:
unexplained variation $= 3353.556$

$(\sum y)^2$

For our current dataset:
$unexplained\ variation = 3353.556$

$$total\ variation = (n-1)(S_y^2) = (\sum y^2) - \frac{(\sum y)^2}{n}$$

either formula works

$total\ variation = 4030.8$

Regression Sum of Squares

$SSR = SSTOTAL - SSE$
$= total\ variation\ in\ y\ minus\ the\ unexplained\ variation\ in\ y$

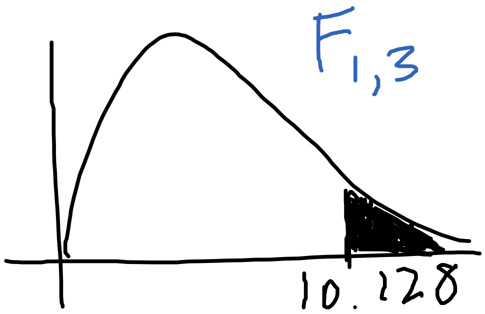
Those 1860s types couldn't resist calling SSR the "explained" variation in Y, or more specifically the variation in y that is explained by the regression relationship.

The ANOVA table for regression

For some reason ANOVA comes up again in regression.

Regression DF stands for the number of x-variables. Since we are not yet in multi-variable regression, Regression DF = 1.

Source	DF	SS	MS	F
Regression	1	677.244	677.244	$\frac{677.244}{1117.852} = 0.6$
Error	3	3353.556	1117.852	
Total	4	4030.8		



$F - test\ statistic = 0.6$

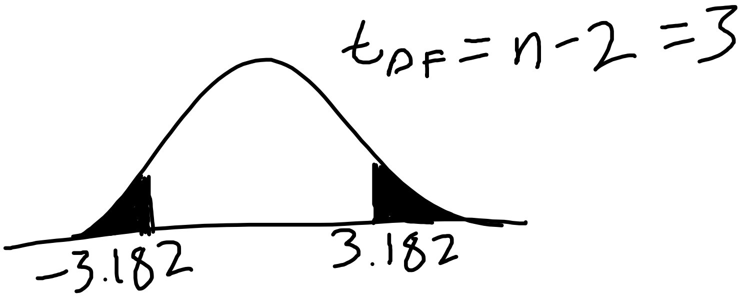
Do not reject the null hypothesis.

H_0: This equation does not give significant predictions for y.
H_A: This equation is significant for predicting y.

Another popular version:

H_0: X and Y are not significantly correlated.
H_A: X and Y are significantly correlated.

$H_0{:}\,rho = 0$
$H_A{:}\,rho \neq 0$



$r = -0.40989939$ (from calculator)

$$test\ statistic = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \approx \frac{-0.4099}{\sqrt{\frac{1-0.4099^2}{3}}} = -0.778360533$$

$F = 10.128$
$\sqrt{10.128} = 3.182$

Oh look t and F are related!

Lecture 14/03/24

March 24, 2014
2:47 PM

t-test for correlation

Test statistic =
$$\frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

$H_A: x\ and\ y\ are\ significantly\ related$

Overall F-test

$H_0: this\ regression\ equation\ is\ not\ significant\ for\ predicting\ y$
$H_a: this\ regression\ equation\ is\ significant\ for\ predicting\ y$

$$test\ statistic = \frac{MSR}{MSE}$$

Warm-up to the longest equation in the course

n = 5

x_i	y_i
12	85
15	92
50	210
40	166
64	312

Try this on your calculators. You should be getting:

$a = 25.61377186 \approx 25.614$
$b = 4.071442766 \approx 4.071$

$\hat{y} = 25.614 + 4.071x$

If you don't know how to use your calculator to find a and b (but not y-hat, that is supposed to be the predicted result), go and ask Kmack for he is the prophet of calculators.

Example:
$\hat{y}(10) \approx 66.328$

For reals:

x_i	y_i	$\hat{y}$	e_i
12	85	74.471	10.529
15	92	86.685	5.319
50	210	229.186	-19.186
40	166	188.471	-22.471

64	312	286.186	25.814
			$\sum e_i = 0.001$

In theory the sum should be zero, but we rounded to three decimal places.

SSE
$= \sum e_i^2$
$= 1678.525 \approx 1678.53$

Since we rounded to three places for the residuals we don't trust the third decimal for the SSE. 1678.53 is our unexplained variation in Y.

Analysis of Variance for Regression

The computer tends to call this ANOVA, but it's really Analysis of Variance for Regression. Make sure your crib sheet is making the distinction.

Source	DF	SS	$MS = \frac{SS}{DF}$	F
Regression	1	33365.47	33365.47	59.633
Error	3	1678.53	559.51	
Total	4	3504.4		

The rule for the regression DF number is the number of x variables. So far we have only done single variable so the DF has been 1.

$$Total\ DF = n - 1 = 5 - 1 = 4$$

$$SSTOTAL = (\sum y^2) - \frac{(\sum y)^2}{n} = (n-1)S_y^2 = 3504.4$$

$$SSR = SSTOTAL - SSE = SSTOTAL \times r^2$$

Why?

$$r^2 = \frac{SSR}{SSTOTAL}$$

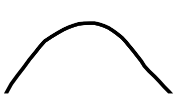
$$SSR = SSTOTAL \times r^2$$

Finding SSR this way is a bit faster and more accurate because you can just get these numbers with buttons on your calculator and the calculator will keep more decimals than you can.

Overall F-test

H_0: *this regression equation is not significant for predicting y*
H_a: *this regression equation is significant for predicting y*

$$test\ statistic = \frac{MSR}{MSE}$$

1  $F_{1,3,0.05}$

Test statistic

$= 59.663 \gg 10.128$

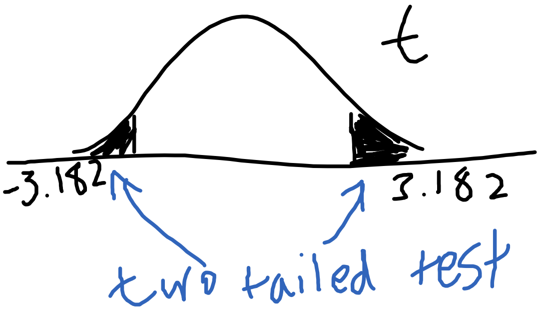
Reject the null hypothesis.

Thus equation is significant for predicting Y.

Kmack admits this test is not very intuitive, as compared to the t-test that tests whether the two variables are co-related whereas the F-test looks like a bunch of calculation steps Kmack tells you to do.

But recall that the two tests are related:

$H_0: X \text{ and } Y \text{ are significantly correlated}$
$H_A: X \text{ and } Y \text{ are significantly correlated}$



The t-test is two tailed because X and Y can be positively correlated or negatively correlated.

Test statistic

$$= \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

For our example,

$$= \frac{0.975757302}{\sqrt{\frac{1 - 0.975757302^2}{5-2}}} \approx 7.72227331$$

Reject null hypothesis X and Y are significantly correlated.

In essence squaring the t-test test statistic and the t-value will give the F-test test statistic and the F-value. These two tests are really the same test, but with different "clothing".

A glimpse to the future: the p-value is coming. Don't forget p-value less than 0.05 means reject!

The Longest Equation in ECON 227

The prediction interval (PI) formula

If $x = x_0$, the point estimate for the value of y that will be observed is $a + bx_0 = \hat{y}(x_o)$.

The interval estimate based on this point estimate is

$$a + bx_0 \pm t\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)}}$$

$$t_{DF} = error\ DF$$

This formula must be written down because no human being can remember it.

For our current data set, find a 95% prediction interval for the value of Y that will be observed if x = 10.

$$25.614 + 4.071(10) \pm (3.182)(\sqrt{559.51})\left(\sqrt{1 + \frac{1}{5} + \frac{(10 - 36.2)^2}{\left(8565 - \frac{181^2}{5}\right)}}\right)$$

$$\cong 66.328 \pm 93.435$$

Terminology

$$\hat{y} = a + bx$$

b

- i) is the regression coefficient (coefficient of x)
- ii) the slope of the line
- iii) the derivative of a+bx
- iv) the marginal contribution of x (according to economists)

In the expression $\hat{y} = a + bx$, if x is increased by 1 unit, we get $a + b(x+1) = a + bx + b$ change in y

b is a statistic that is calculated from a sample.

$b \pm tS_b$ → S_b is the so-called std. error of b

$$= b \pm t\left(\sqrt{\frac{MSE}{\left((\sum x^2) - \frac{(\sum x)^2}{n}\right)}}\right)$$

More on this next class.

Lecture 14/03/26

March 26, 2014
2:38 PM

Announcements

- New assignment up on MyCourses along with two past exams. Today will likely be the last day of new material.
- No new material on election day (April 7th), class will still be held as a review session.

Recap

$\hat{y} = a + bx$

If we have $x + 1$ instead of x, we get $a + b(x + 1) = a + bx + b$. The change in $\hat{y}$ when x increases by one unit is b, the coefficient of x. We call the coefficient of x the *marginal contribution* of x.

The shortest equation in the course (again)
CI for the marginal contribution x is:
$b \pm tS_b$

$$S_b = \sqrt{\frac{MSE}{(\sum x^2) - \frac{(\sum x)^2}{n}}}$$

Example: 2009 Supplemental Q2 (will be on MyCourses)

Q2 a) Regress the price of gold on the price of aluminum. (The data is the same as in Q1)

Translation: put the data into your calculator and write down the $\hat{y} = a + bx$ equation. As always, practice with your calculator and if you don't understand go ask Kmack.

The terminology is always this:
"Regress y on x".

So for "regress the price of gold on the price of aluminum" means y represents the price of gold and x represents the price of aluminum.

After punching the data into your calculator you should get:

$a = 131.3396$
$b = 3.3175$
$n = 10$

So, $\widehat{gold} = 131.3396 + 3.3175(aluminum)$

The partial calculations were given in the question which you can use to check if you entered the data in the calculator.

Q2 b) Fill out the Analysis of Variance table for the regression.

Remember this is not ANOVA! Don't use ANOVA formulas here.

Source	DF

Regression	1
Error	8
Total	9

- Regression DF is the number of x variables.
- Total DF is $n-1$.
- Regression DF + Error DF must equal the total.

Simplest way to fill out the table

Source	DF	SS	MS	F
Regression	1	32023.978	32023.978	2.811584004
Error	8	91120.103	11390.013	
Total	9	123144.081		

$$SSTOTAL = \left(\sum y^2\right) - \frac{\left(\sum y\right)^2}{n} = (n-1)S_y^2$$
$$SSR = SSTOTAL \times r^2$$
$$R^2 = \frac{SSR}{SSTOTAL}$$

In simple regression small r and big R are the same. Big R does not exist for multiple regression.

MS Regression (MSR) and MS Error (MSE) are the SSR and SSE values divided by their corresponding DF values.

The F-test statistic is $\frac{MS\ Regression}{MS\ Error}$

We will use the numbers in the highlighted cells later.

Q2 c) Form a 95% prediction interval (PI) for the price of gold when the price of aluminum is 75 cents.

Make sure you watch out for units. In our case, we got the data in cents so we will stick to cents.

$$a + bx_0 \pm t\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)}}$$

For the t, use DF = error DF = 8 for our example. Make sure you are choosing $\alpha = 0.025$ keeping in mind this is a two-tailed test.

$$380.16 \pm 2.306\left(\sqrt{11390.013}\right)\left(\sqrt{1 + \frac{1}{10} + \frac{(75 - 76.21)^2}{\left(60989.29 - \frac{762.1^2}{10}\right)}}\right)$$

$$380.16 \pm 258.18$$

The error is big because we have only 10 data points.

Q2 d) Estimate the standard error of the regression coefficient

Regression coefficient is b, the standard error is S_b.

$$S_b = \sqrt{\frac{11390.013}{60989.29 - \frac{762.1^2}{10}}} = 1.9785$$

Supplementing question: Give 95% CI for the marginal contribution of x.

$b \pm tS_b = 3.3175 \pm 2.306(1.9785)$
3.3175 ± 4.5624

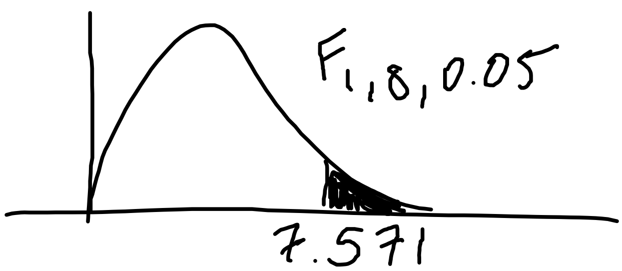
Q2 e) What proportion of the variation in the price of gold is explained by the regression relationship?

$$\frac{SSR}{SSTOTAL} = \frac{32023.978}{123144.081} = 0.26005 \approx 26\%$$

Or (for simple regression only):

$r^2 = 0.50995^2 = 0.26005 \approx 26\%$

We didn't check if the model is significant for predicting x. We can check this easily with the F-test statistic.



Test statistic = 2.81
Do not reject the null hypothesis. This equation is not significant for predicting the price of gold.

Example: Multiple regression, 2009 Supplemental Q1

The sum of squares never changes in regression, you can reuse it if it was given in another question for a simple regression.

Q1 a) Predict the price of gold if silver is $10 per pound and aluminum is 80 cents a pound.

Since the price of copper is not given, we use MODEL II. There was a typo, silver is in $ per ounce.

$\widehat{gold} = -89.26 + 20.556(10) + 3.7253(80) \cong \$414/ounce$

Q1 d) If the price of copper and silver do not change, give a 95% CI for the change in price of gold.

Translation: Ceteris paribus. This is a question about the marginal contribution of silver.

Since the price of copper is now known in this question, use MODEL I.

A point estimate for the marginal contribution

$= 20.358$

Reason:
$-95.67 + 0.2619(copper) + 20.358(silver + 1) + 3.5317(aluminum)$
$= -95.67 + 0.2619(Cu) + 20.358(Ag) + 20.358 + 3.5317$

CI:
$b_{Ag} \pm tS_{b_{Ag}}$

DF = error DF = 6

$= 20.358 \pm 2.447(2.722)$
$= 20.358 \pm 6.661$

Statement: If the price of silver increases by $1/ounce, the predicted price of gold will change by somewhere between $13.70 and $27.07 with a confidence level of 95%.

We are not done, but we will finish next class and begin review with the 2012 final exam.

Lecture 14/03/31

March 31, 2014
2:38 PM

Announcements

New test used in recent textbooks is acceptable:

- If one of the two variances are more than twice as big as the other, then the two are significantly different.

Correction of example: Multiple regression, 2009 Supplemental Q1

(The correction is highlighted)

The sum of squares never changes in regression, you can reuse it if it was given in another question for a simple regression.

Q1 d) If the price of copper and silver do not change, give a 90% CI for the change in price of gold.

CI:
$b_{Ag} \pm tS_{b_{Ag}}$

DF = error DF = 6

$= 20.358 \pm 1.943(2.722)$
$= 20.358 \pm 5.289$

Statement: If the price of silver increases by $1/ounce, the predicted price of gold will change by somewhere between $15.07 and $25.65 with a confidence level of 95%.

Continuing 2009 Exam

Q1 b)

MODEL I $R^2 = 0.932$
MODEL II $R^2 = 0.930$

If a variable is deleted or omitted from a multiple regression model, R^2 decreases, except in some pathological situations. On the other hand, if an extra x variable is included, R^2 increases.

For example, if R^2 is already 100%, there's no point in trying to add another x variable.

Even though it's obvious R^2 dropped from MODEL I to MODEL II, answers must always include some statistical language. In this case we can use the p-value.

Copper is the variable that caused the change because it is the one not in MODEL II, but is in MODEL I.

In the 2009 exam question 1b), the null hypothesis is the reduction in R^2 is not significant if copper is omitted and the alternative hypothesis is that the reduction is significant.

$H_0: reduction\ in\ R^2\ is\ not\ significant$
$H_A: reduction\ in\ R^2\ is\ significant$

This is called the individual t-test.

The individual t-test for copper:

$p - value = 0.702$

We don't have a significant decrease because the p-value is greater than 0.05. We do not reject the null hypothesis.

Alternative method:

Test statistic

$$= \frac{b_{copper}}{S_{b_{copper}}} = \frac{coefficient}{SE\ coefficient} = \frac{0.2619}{0.6514} = 0.40$$

The error DF = 6 (taken from MODEL I because that is where we go the numbers for copper). The critical t-value is 1.943 (from t-table).

The p-value method is the easier than this but you can still do it if you really like null and alternative hypothesis.

Q1 c) If the variable "silver" had been omitted instead of "copper" would the reduction in R^2 be greater or less?

Since the t-test p-value is listed as 0.000 (on the computer printout on the exam), the reduction would have been significant.

The reduction in R^2 would have been greater.

Q1 e)

This is just terminology. Standard error of the estimate:

$$= \sqrt{MSE} = \sqrt{1232} \approx 35$$

No questions on final on r-sqr-adj (R-Squared-Adjusted).

Review 14/03/31

March 31, 2014
2:41 PM

McGill

APRIL 20^{th}, 2012

Final Examination

ECONOMIC STATISTICS
ECON 227
April 20^{th}
14:00 to 17:00

Examiner: K.MacKenzie **Associate Examiner: J.Kurien**

INSTRUCTIONS:

- **This is a CLOSED BOOK examination**
- **One legal-sized CRIB SHEET permitted**
- **You are permitted dictionaries**
- **CALCULATORS of any type are permitted**
- **This examination is PRINTED ON BOTH SIDES of the paper**
- **Do all seven questions**
- **Each question is worth the same amount**
- **Answer on examination booklets provided**
- **This examination has 7 pages, including the cover**

Tips:

- Most people make the mistake of using the wrong test. Be careful if your test is for proportions or otherwise.

2. a) **Regress the price of aluminum on the price of copper. Use either the statistics buttons on your calculator or the partial computations shown:**

$$\sum Copper = 877.5 \quad \sum Alum = 762.1 \quad \sum Copper^2 = 81936.25$$
$$\sum Copper \times Alum = 68981.11 \quad \sum Alum^2 = 60989.29$$

b) **Calculate a 90% prediction interval for the price of aluminum when copper sells for 75 cents a pound.**

c) **What proportion of the variation in the price of aluminum is explained by the regression relationship with the price of copper?**

d) **Form a 90% confidence interval for the marginal contribution of copper.**

e) **Test at the 10% level whether the coefficient of determination (i.e. R^2) is significantly less than the R^2 calculated in MODEL III.**

3. **The table below gives the holdings in Eurelian bonds of a randomly-selected group of Eurelians in different fields.**

	Blue Collar	White Collar	Entertainment	Politics	Total
East Eurelia	12000	15000	9000	12500	48500
Central Eurelia	13500	16200	10800	14000	54500
West Eurelia	12500	16000	9600	12000	50100
Total	38000	47200	29400	38500	153100

✓ ANOVA test   If this were a chi-square test, the word "depend" would be in the question.

a) **Test whether the mean holdings are significantly different in the different fields after the effect of the region (East, Central, West) has been removed.**

b) **Form a 90% CI (confidence interval) for the difference in mean holdings between Blue and White collar Eurelians.**

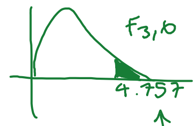
3. a)

H_0: the means are not significantly different

$(\mu_1 = \mu_2 = \mu_3 = \mu_4)$

H_A: at least one is significantly different

This is two way ANOVA because we are removing the effects of regions



Source	DF	SS	MS	F
Treatments	3	52849166.67	17616388.89	MSTR/MSE = 104.3
Blocks	2	4826666.67	2413333.33	
Error	6	1013333.33	168888.89	
Total	11	58689166.67		

$n = 12$

$SSTOTAL = (\sum x^2) - \frac{(\sum x)^2}{n}$

$= 2011990000 - \frac{153100^2}{12}$

$SSTR = \frac{38000^2}{3} + \frac{47200^2}{3} + \frac{29400^2}{3} + \frac{38500^2}{3} - \frac{153100^2}{12}$

$SSB = \frac{48500^2}{4} + \frac{54500^2}{4} + \frac{50100^2}{4} - \frac{153100^2}{12}$

ECON 227

$x_1 - x_2 \pm t\sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$

DF=6

90% CI, t=1.943

$= \frac{38000}{3} - \frac{47200}{3} \pm (1.943)\sqrt{\frac{168888.88}{3} + \frac{168888.88}{3}}$

$= -3066.67 \pm 651.97$

7. Some summarized data on hours of sleep are given for young urban professionals (yuppies) and for geriatric urban professionals (guppies).

	Young Urban Professionals	Geriatric Urban Professionals
Sample size	16	12
Mean hours of sleep	7.2	6.6
Standard deviation	2.4	1.6

a) Test whether the variances are significantly different for yuppies and guppies.

b) Test whether the mean numbers of hours of sleep is significantly higher for yuppies than for guppies.

c) Estimate the standard error of the difference of the sample means.

d) In the sample from which the data above are taken, there were 23 subjects who could not be classified either as yuppies or guppies. The manufacturer has a preconceived notion that yuppies and guppies both are around one-third of the population. Do the data indicate that the proportions are significantly different amongst the three groups: yuppie, guppie, and other? Test using $\alpha = 0.15$.

2009 Final ECON 227

April 2, 2014
12:22 PM

ECON 227

1. **The Eurelian Bureau of Mines produces data on the price of minerals. Shown here are prices for minerals over ten randomly-selected months. Two multiple-regression models have been calculated by MINITAB. Use numbers from the output to answer Question 1.**

Gold (\$/*ounce*)	*Copper* (*cents*/*pound*)	*Silver* (\$/*ounce*)	*Aluminum* (*cents*/*pound*)
161.1	64.2	4.4	39.8
308.0	93.3	11.1	61.0
613.0	101.3	20.6	71.6
460.0	84.2	10.5	76.0
376.0	72.8	8.0	76.0
424.0	76.5	11.4	77.8
361.0	66.8	8.1	81.0
318.0	67.0	6.1	81.0
438.0	120.5	6.5	110.1
382.6	130.9	5.5	87.8

MODEL I

Regression Analysis: Gold versus Copper, Silver, Aluminum

```
The regression equation is
Gold = - 95.67 + 0.2619 Copper + 20.358 Silver + 3.5317 Aluminum

Predictor    Coef   SE Coef      T      P
Constant   -95.67     63.35  -1.51  0.182
Copper     0.2619    0.6514   0.40  0.702
Silver     20.358     2.722   7.48  0.000
Aluminum   3.5317    0.8459   4.18  0.006

S = 37.4095   R-Sq = 93.2%   R-Sq(adj) = 89.8%

Analysis of Variance

Source          DF      SS     MS      F      P
Regression       3  114747  38249  27.33  0.001
Residual Error   6    8397   1399
Total            9  123144
```

ECON 227

MODEL II

Regression Analysis: Gold versus Silver, Aluminum

```
The regression equation is
Gold = - 89.26 + 20.556 Silver + 3.7253 Aluminum


Predictor     Coef  SE Coef      T      P
Constant    -89.26    57.52  -1.55  0.165
Silver      20.556    2.512   8.18  0.000
Aluminum    3.7253   0.6526   5.71  0.001


S = 35.0980   R-Sq = 93.0%   R-Sq(adj) = 91.0%


Analysis of Variance

Source          DF      SS     MS      F      P
Regression       2  114521  57261  46.48  0.000
Residual Error   7    8623   1232
Total            9  123144
```

a) **Predict the price of gold for a month when silver has a price of $10 per pound and aluminum has a price of 80 cents per pound.**

b) **Is the reduction in R^2 from MODEL I to MODEL II statistically significant? Justify your answer from the output.**

c) **If the variable 'silver' had been deleted from the model instead of 'copper', would the reduction in R^2 have been greater or less?**

d) **If copper and aluminum prices do not change, but the price of silver increases by $1 per ounce, give a 90% confidence interval for the change in the price of gold. Use MODEL I.**

e) **What is the standard error of the estimate in MODEL II?**

Review 14/04/02

April 2, 2014
2:39 PM

==Something relevant to the assignment==

Homework question 1
a) We can switch to the very easy test to see whether two variances or standard deviations are significantly different

Rule: the standard deviations (or variances, one implies the other) are significantly different if and only if one of the sta ndard deviations is more than double the other one.

b)
$H_0: \sigma \leq \sigma_0$
$H_A: \sigma > \sigma_0$

Test statistic
$$= \frac{(n-1){S_x}^2}{{\sigma_0}^2}$$
Only in this instance will we use this test statistic, otherwise use Pearson's.

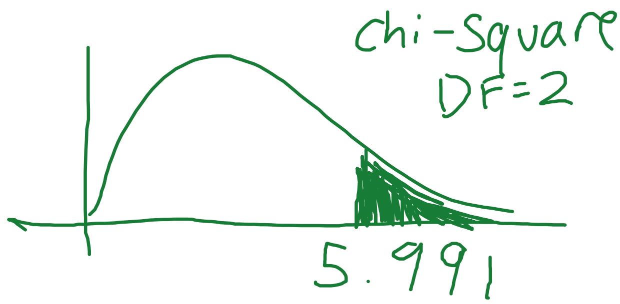
Continuing the 2012 Final

5

4. a) A large group of Eurelians has been polled concerning political preferences. It is suspected that there is virtually a tie amongst the three major political parties in Eurelia. Here are the results of the poll:

Atavist Party preferred	Recidivist Party preferred	Party Party preferred
300	340	360

Test whether there is a significant difference in party preferences in Eurelia.



4. a) This is like the summer drinks example.

$$H_0: P_{ATA} = P_{REC} = P_{PARTY} = \frac{1}{3}$$
$$H_A: at\ least\ one\ P_j \neq \frac{1}{3}$$

DF = 2 because we can only pick two proportions before the last one can only be one value.

O_i	E_i
300	333 1/3
340	333 1/3
360	333 1/3
1000	

Test statistic

$$= \frac{(300-333.333)^2}{333.333} + \frac{(340-333.333)^2}{333.333} + \frac{(360-333.333)^2}{333.33}$$

OR

$$= \frac{(300)^2}{333.333} + \frac{(340)^2}{333.333} + \frac{(360)^2}{333.33} - 1000$$
$$= 5.6$$

Do not reject H_0. There is no significant difference in the supports for the three parties

b) Another hypothesis test is performed on the data in part a). Let p stand for the proportion of Atavist Party supporters in the Eurelian population. The alternative hypothesis for the test is $H_a: p \neq \frac{1}{3}$. What is the p-value for this test?

4. b)

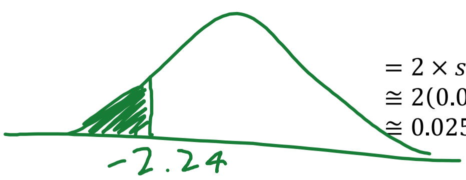
Let p=proportion of Atavist Party supporters

$$H_0: p = \frac{1}{3}$$
$$H_a: p \neq \frac{1}{3}$$

This is a two-tailed test, remember to double the area at the end!

Test statistic

$$= \frac{p-p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.3-\frac{1}{3}}{\sqrt{\frac{\frac{1}{3}(1-\frac{1}{3})}{1000}}} = -2.24$$



$$= 2 \times small\ area$$
$$\cong 2(0.0125)$$
$$\cong 0.025$$

c) I have a textbook in which the value listed in the t tables for 150 degrees of freedom in the column headed 0.025 is 1.976. If I look in the tables of the F distribution with numerator df equal to 1 and denominator df equal to 150, what number should appear on the 0.05 page?

4. c)
$t_{150,0.025} = 1.976$
If I look in the F-tables, $F_{1,150,0.05}$, what number will I see there?
$1.976^2 = 3.905$, because the F-test statistic is the t-test statistic squared.

The reason why we use 0.05 for F and not 0.025 is because when you square the negative side of the t graph, it will also come to the positive side so there a two-tailed t-test at 0.025 corresponds to the right tailed F-test at 0.05

5. **In Eurelia there are three categories of bicycle license, with different fees for each. The table below is based on a random sample of 262 Eurelian bicycle-license holders.**

	License Fee		
	$25	$30	$27
Female	28	32	24
Male	36	44	40
Other	18	24	16

ECON 227

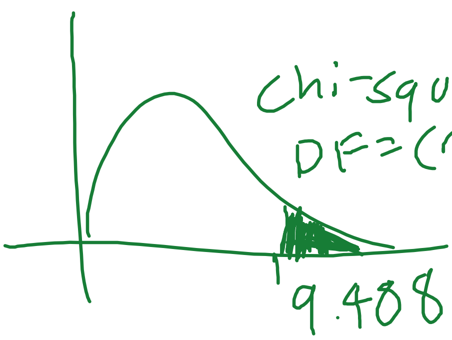
Chi-Square

6

a) **Does the type of license bought depend significantly on whether the Eurelian is Female, Male, or Other?**

b) **How large a sample would be needed in order to establish a 90% confidence interval for the mean amount paid by Eurelians in the Other category?**

H_0: *License types and gender are independent*
H_A: *License types and gender are dependent*



chi-square
$DF=(r-1)(c-1)=4$

	25$	30$	27$	Total
Female	28	32	24	84
Male	36	44	40	120
Other	18	24	16	58
Total	82	100	80	262

E_{ij} *table*

(84 x 82)/262 = 26.2901	(84 x 100)/262 = 32.0611	(84 x 80)/262 = 25.6489	84
(120 x 82)/262 = 37.5573	45.80	36.64	120
18.15	22.14	17.71	58
82	100	80	262

$$\sum \frac{(O-E)^2}{E} = \frac{(28-26.29)^2}{26.29} + \cdots + \frac{(16-17.71)^2}{17.71}$$
$$= 0.98 \ (which\ is\ less\ than\ 9.488)$$

Do not reject the null hypothesis. License type and gender are independent (not significantly dependent)

Make sure you know how to do frequencies on your calculator (usually it's add a semicolon, then the frequency)

5. b) Pretend margin of error is 50 cents

$$\frac{1.645^2 \times 2.1588^2}{(0.50)^2}$$

7

7. Some summarized data on hours of sleep are given for young urban professionals (yuppies) and for geriatric urban professionals (guppies).

	Young Urban Professionals	Geriatric Urban Professionals
Sample size	**16**	**12**
Mean hours of sleep	**7.2**	**6.6**
Standard deviation	**2.4**	**1.6**

a) Test whether the variances are significantly different for yuppies and guppies.

b) Test whether the mean numbers of hours of sleep is significantly higher for yuppies than for guppies.

7. a)

Since neither standard deviation is more than double the other, they are not significantly different. Neither the standard deviation nor the variance.

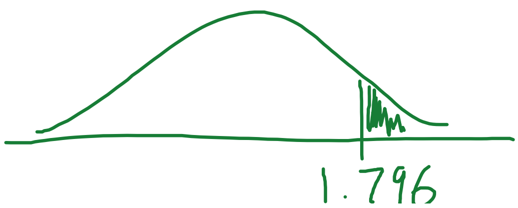
7. b)

Lazy DF (acceptable but faster method) is the smaller of $n_2 - 1$ and $n_1 - 1$. In this question it is 11.

$H_0: \mu_y \le \mu_G$
$H_a: \mu_y > \mu_G$

Test statistic

$$= \frac{7.2-6.6}{\sqrt{\frac{2.4^2}{16}+\frac{1.6^2}{12}}} = 0.79$$



Do not reject H_0. The mean is not significantly greater.

Review 14/04/07

April 7, 2014
2:42 PM

Announcements

- Please go to Kmack's office hours instead of just anytime because he is behind on paperwork
- Today's review is the April 2011 exam
- Bagpipes on Wednesday

6. Independent random samples of university students were selected from MacMillan University and L'Université d'Eurélie. Here are some of the data that were collected on daily coffee consumption.

	MU	U d'E
sample size	8	12
mean coffee consumption	560 ml	625 ml
standard deviation	68 ml	22 ml

c) Form a 95% CI (confidence interval) for the difference in mean coffee consumption.

$$DF = \min(8,12) - 1$$

$$\bar{x}_1 - \bar{x}_2 \pm t\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$= 560 - 625 \pm (2.365)\sqrt{\frac{68^2}{8} + \frac{22^2}{12}}$$

$$= -65 \pm 58.8$$

d) Estimate the standard error of the difference of the sample means.

$$= \sqrt{\frac{68^2}{8} + \frac{22^2}{12}}$$

$$= 24.866$$

1. The data for the first two questions give the ice-cream consumption over 12 four-week periods from March 1950 to April 1951

t	**The four-week-period number**
consump	**The ice-cream consumption (in ml per capita)**
price	**The price of ice-cream (in dollars)**
income	**The weekly family income (in dollars)**
temp	**The mean ambient temperature (in degrees)**

t	consump	price	income	temp
1	193	0.27	78	5
2	187	0.282	79	13
3	196.5	0.277	81	17
4	212.5	0.28	80	20

2	187	0.282	79	13
3	196.5	0.277	81	17
4	212.5	0.28	80	20
5	203	0.272	76	21
6	172	0.262	78	18
7	163.5	0.275	82	16
8	144	0.267	79	8
9	134.5	0.265	76	0
10	128	0.277	79	-4
11	143	0.286	82	-2
12	149	0.27	85	-3

MODEL I

```
The regression equation is
consump = - 47.6 - 4.751 t + 370.3 price + 1.678 income + 1.3704 temp

Predictor     Coef   SE Coef      T      P
Constant     -47.6     182.0  -0.26  0.801
t           -4.751     1.771  -2.68  0.031
price        370.3     674.8   0.55  0.600
income       1.678     2.017   0.83  0.433
temp        1.3704    0.6125   2.24  0.060

R² = 84.7%


Analysis of Variance

Source          DF      SS      MS     F      P
Regression       4  7892.6  1973.2  9.71  0.006
Residual Error   7  1423.1   203.3
Total           11  9315.7
```

MODEL II

```
The regression equation is
consump = - 56.0 + 878.4 price - 0.473 income + 2.4339 temp


Predictor     Coef  SE Coef      T      P
Constant     -56.0    242.4  -0.23  0.823
price        878.4    862.8   1.02  0.338
income      -0.473    2.466  -0.19  0.853
temp        2.4339   0.6220   3.91  0.004

Analysis of Variance

Source          DF      SS      MS     F      P
Regression       3  6429.3  2143.1  5.94  0.020
Residual Error   8  2886.4   360.8
Total           11  9315.7
```

a) Use MODEL I to predict the ice-cream consumption when t = 14, the price is 35 cents, the family income is 80 dollars per week, and the ambient temperature is 5 degrees.

Make sure you are using the right units.

$\widehat{consump} = -47.6 - 4.751(14) + 370.3(0.35) + 1.678(80) + 1.3704(5)$

$\widehat{consump} = 156.583$

b) Is the reduction in R^2 from MODEL I to MODEL II statistically significant? Justify your answer numerically.

Individual t-test p-value $= 0.031 < 0.05$. The change in R^2 is significant.

c) Construct a 95% confidence interval for the marginal contribution of the temperature.

It appears in both models, so in this situation either answer is correct. But we will use both models here for the sake of demonstration.

MODEL I

$Error\ DF = 7$

$1.3704 \pm 2.365(S_b)$

$S_b\ is\ from\ the\ printout$

$= 1.3704 \pm 2.365(0.6125)$

$= 1.37 \pm 1.44$

DO NOT USE

$$S_b = \sqrt{\frac{MSE}{blah\ blah}}$$

For multiple regression. This formula works ONLY for single regression.

MODEL II

$2.4339 \pm 2.306(0.6220)$

$= 2.43 \pm 1.43$

d) Is MODEL II significant for predicting ice-cream consumption? Justify your answer numerically.

The F-test p-value is 0.02 < 0.05. Therefore the model is significant.

NB: Remember that p-value < 0.05 means reject the null hypothesis and the model **is** significant.

2. Use the data of question 1.

a) Regress the ice-cream consumption on the price. Either use statistics buttons on your calculator or formulas. The following partial calculations are provided.

$$\sum x^2 = 0.898745 \quad \sum y^2 = 351372$$

$$\sum x = 3.283 \quad \sum y = 2026 \quad \sum xy = 554.6915$$

NB: remember to regress y on x ALWAYS.

$y = ice - cream$

$x = price$

$\hat{y} = -28.437 + 721.063x$

b) Test whether the price is significant for predicting consumption.

Source	DF	SS	MS	F
Regression	1	296.837	296.837	0.329
Error	10	9018.830	901.883	
Total	11	9315.667		

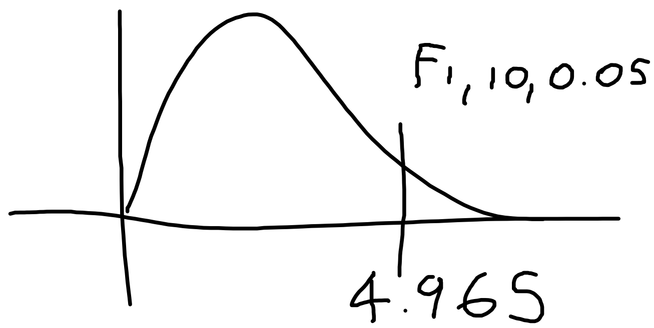
$$SSTOTAL = \sum y^2 - \frac{(\sum y)^2}{n} = 351372 - \frac{202.6^2}{12} = 9315.667$$

NB. $SSTOTAL = (n-1)S_y^2$

$$SSR = SSTOTAL \times r^2 = 296.837$$

H_0: *the model is not significant for predicting consumption*
H_A: *the model is significant for predicting consumption*

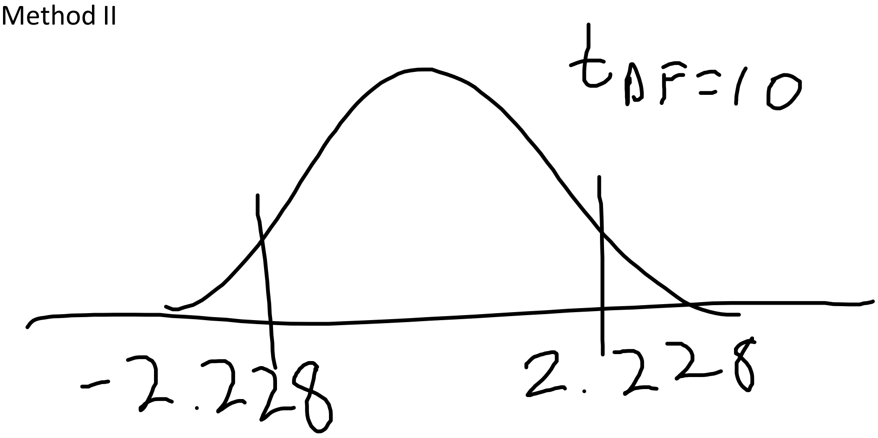
Method I



Test statistic = 0.329

Do not reject H_0. The model is not significant for predicting consumption.

Method II



Test statistic

$$= \frac{r}{\sqrt{\frac{1-r^2}{n-1}}}$$

$$= \frac{0.1785}{\sqrt{\frac{1-0.1785^2}{10}}} = 0.573$$

Do not reject H_0. The model is not significant.

Again, the F-test statistic is the square of the t-test statistic.

c) Calculate a 90% prediction interval for the consumption of ice cream when the price is 35 cents.

$$= a + bx_0 \pm t\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)}}$$

$$= 223.93 \pm (1.812)(901.883)(more\ stuff)$$

$$= 223.93 \pm 183.02$$

d) What proportion of the variation in ice-cream consumption is explained by the regression relationship with the price?

$$r^2 = 0.03186$$

So aboot 3.186%

e) Calculate the estimate of the standard error of the regression coefficient.

$$\sqrt{\frac{MSE}{\sum x^2 - \frac{(\sum x)^2}{n}}} = \sqrt{\frac{901.883}{0.898745 - \frac{3.283^2}{12}}} = 1256.865$$

Review 14/04/09

April 9, 2014
2:32 PM

2011 Final continued

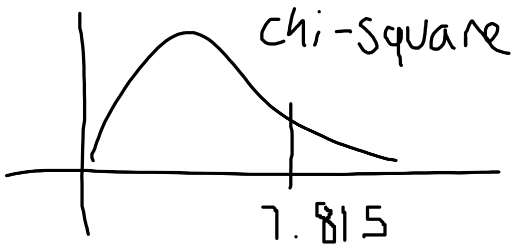
Answers to 2005 Final are on MyCourses.

3. a)

There are four political parties in Eurelia: Recidivist, Atavist, Ergodist, and Anachronist. A pollster desires to learn if the political party supported depends significantly on whether the voter is urban or rural. The following preliminary data have been collected. What conclusion will the pollster draw?

	Recidivist	Atavist	Ergodist	Anachronist
Rural	10	22	20	26
Urban	20	28	40	34

Key word is "depends" so this is a chi-squared independence test.



H_0: *party and domicile are independent*
H_A: *party and domicile are significantly dependent*

$$Chi-square\ DF = (r-1)(c-1) = (2-1)(4-1) = 3$$

$$E_{ij} = \frac{r_{total} \times c_{total}}{n}$$

					Total
	11.7	19.5	23.4	23.4	78
	18.3	30.5	36.6	36.6	122
Total	30	50	60	60	200

Test statistic

$$= \frac{(10-11.7)^2}{11.7} + \frac{(22-19.5)^2}{19.5} + \cdots + \frac{(34-36.6)^2}{36.6} = 2.21$$

Alternative method

$$= \left(\sum \frac{O^2}{E}\right) - n$$

$$= \frac{10^2}{11.7} + \frac{22^2}{19.5} + \cdots + \frac{34^2}{36.6} - 200 = 2.21$$

Reject H_0. Party preferred and domicile are not significantly dependent.

b) In the last election the Recidivists got 22% of the vote, the Atavists got 30%, the Ergodists got 35%, and the Anachronists got 13%. Test whether the proportions are now significantly different from those in the last election, based on the poll results above.

Like the Professor X example.

NB. The morning section has this wrong.

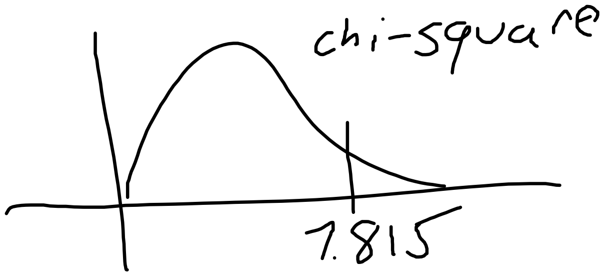
$H_0: p_{REC} = 0.22, p_{AT} = 0.30, p_{ERG} = 0.35, p_{AN} = 0.13$
H_A: *at least one is sigificantly different from the last election*

O	E

chi-square

H_A: *at least one is sigificantly different from the last election*

O	E
30	44
50	60
60	70
60	26
200	200



NB. $k = \#\ of\ categories$ ("kategories")
$DF = k - 1 = 3$

Test statistic

$$= \frac{(30-44)^2}{44} + \frac{(50-60)^2}{60} + \frac{(60-70)^2}{70} + \frac{(60-26)^2}{26} = 52.01$$

NB. Could've began with the (60-26) term and we would already be in the rejection region.

Reject H_0. At least one of the results is significantly different.

4. **In Eurelia three of the four political parties agreed to the following study. Small independent random samples of Eurelian voters were selected, and the table below was compiled.**

Some initial computations have been made.

For the Atavist voters in the table the sum of their incomes is 186, and the sum of the squares of their incomes is 7474.

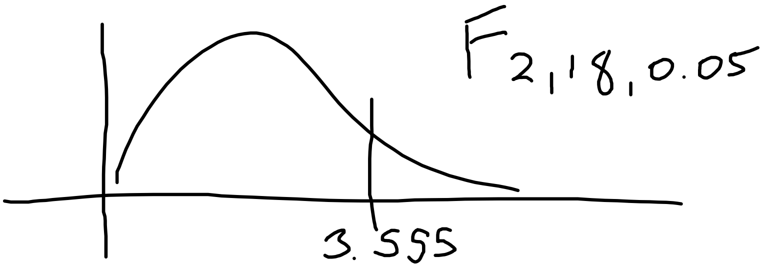
For the Recidivist voters in the table the sum of their incomes is 277, and the sum of the squares of their incomes is 11327.

For the Anachronist voters in the table the sum of their incomes is 371, and the sum of the squares of their incomes is 16601.

Party Affiliation	Income ($1000s)	Age	Party Affiliation	Income ($1000s)	Age
Atavist	23	26	Recidivist	40	39
Atavist	34	33	Anachronist	47	50
Atavist	32	35	Anachronist	45	44
Atavist	54	53	Anachronist	38	41
Atavist	43	46	Anachronist	46	45
Recidivist	35	34	Anachronist	43	46
Recidivist	33	36	Anachronist	28	27
Recidivist	55	54	Anachronist	34	37
Recidivist	44	47	Anachronist	23	22
Recidivist	36	35	Anachronist	67	70
Recidivist	34	37			

a) Test whether there is a significant difference in the mean incomes amongst voters of the three political parties.

$H_0: \mu_1 = \mu_2 = \mu_3$
$H_A: at\ least\ one\ is\ significantly\ different.$



	Total
Atavist	186
Recidivist	277
Anachronist	371

$$SSTOTAL = \sum x^2 - \frac{(\sum x)^2}{n}$$
$$SSTOTAL = 35402 - \frac{834^2}{21}$$

Source	DF	SS	MS	F
Treatments	2	52.2159	26.11	0.21
Error	18	2228.0698	123.78	
Total	20	2280.2857		

$$SSTR = \frac{186^2}{5} + \frac{277^2}{7} + \frac{371^2}{9} + \frac{834^2}{21} = 52.2159$$

Do not reject H_0. There is not a significant difference amongst the means

b) Give the theoretical requirements for the test you performed in a)

Everything as usual and also that variances are the same (i.e. homoscedascity)

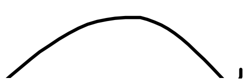
c) Construct a 90% CI (confidence interval) for the mean income of Recidivist-Party affiliates.

$t_{DF=error\ DF=18}$

$$\frac{277}{7} \pm 1.734\left(\frac{\sqrt{123.78}}{\sqrt{7}}\right) = 39.571482 \pm 7.29$$

d) Test whether income and age are significantly correlated.

$H_0: rho = 0$
$H_A: rho \neq 0$



$t_{DF=n-2=19}$

Test statistic

$$= \frac{r}{\sqrt{\frac{1-r^2}{n-1}}}$$

$$= \frac{0.9822}{\sqrt{\frac{1-0.9822^2}{20}}} = 22.79$$

5. A government commission in Eurelia has collected some data on job satisfaction in independent random samples of workers on fixed hours (500 workers) *versus* flex time (300 workers).

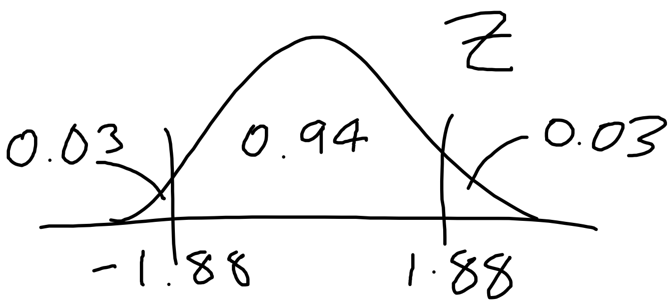
	Fixed Hours	Flex Time
Satisfied	245	105
Not Satisfied	255	195

a) How large a sample is needed to form a 94% confidence interval for the proportion of satisfied workers amongst those on fixed hours? The desired margin of error is 2 percentage points.

$$\frac{Z^2\bar{p}(1-\bar{p})}{E^2}$$

$$= \frac{(1.88)^2\left(\frac{245}{500}\right)(0.51)}{(0.02^2)}$$

$$= 2208.1 \approx 2209$$



b) Use the p-value method to test whether the proportion of satisfied workers amongst those on flex time significantly exceeds the proportion of satisfied workers amongst those on fixed hours.

$H_0: p_{FL} \leq p_{FX}$
$H_a: p_{FL} > p_{FX}$

Test statistic

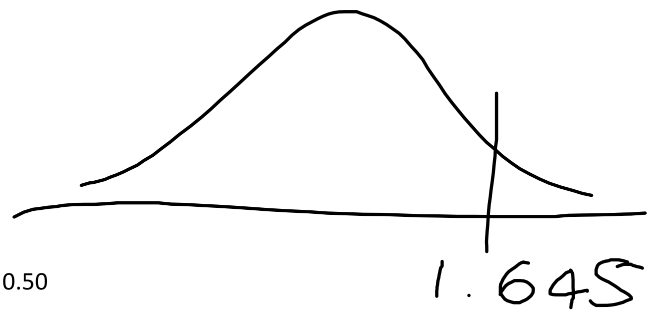
$$= \frac{0.35 - 0.49}{\sqrt{\frac{(0.35)(0.65)}{300} + \frac{(0.49)(0.51)}{500}}} = -3.94$$

$p - value = P(Z > 3.94) \cong 1$

Do not reject H_0. The proportion on FL is not significantly greater than FIX.

c) Test whether the whether the proportion of satisfied workers amongst those on flex time significantly exceeds 50%.

$H_0: p_{FL} \leq 0.50$
$H_A: p_{FL} > 0.50$



Test statistic

$$= \frac{0.35 - 0.50}{\sqrt{\frac{(0.50)(0.50)}{300}}} = -15.996$$

Do not reject H_0. It is not significantly greater than 0.50