MAT 2379A 04/09/2013

Ch. 1: Interpeting Probabilities

A random experiment has an outcome determined by chance and cannot be predicted

The set of all possible outcomes of a random experiment (S) is called the sample space.

An event is a collection of possible outcomes. Example: flipping a coin has 2 possible outcomes, head (H) or tail (T). $S = \{H, T\}$. An event is: $A = \{$a coin lands on heads$\} = \{H\}$

A probability of an event is a number between 0 and 1. (this can also be expressed as a percentage) Example: $P(A) = 0.5 = 1/2 = 50\%$, this is the probability of heads (if the coin is fair)

There are 3 methods for computing probabilities:

1) The personal method
   - one shot situations
   - not accurate
   - subjective
   - ex = will a person pass a driver's test
2) The related frequency method
   - repeatable experiments
   - experiment is ran n times
   - $n(E)$ = number of times our event E occured
   - $P(E) = \frac{n(E)}{n}$
   - ex = treatment of a disease where 100 patients were treated
     $E = \{$treatment was succesfull$\}$
     $n = 100$

$n(E) = 77$

$P(E) = \frac{n(E)}{n} = \frac{77}{100} = 0.77 = 77\%$

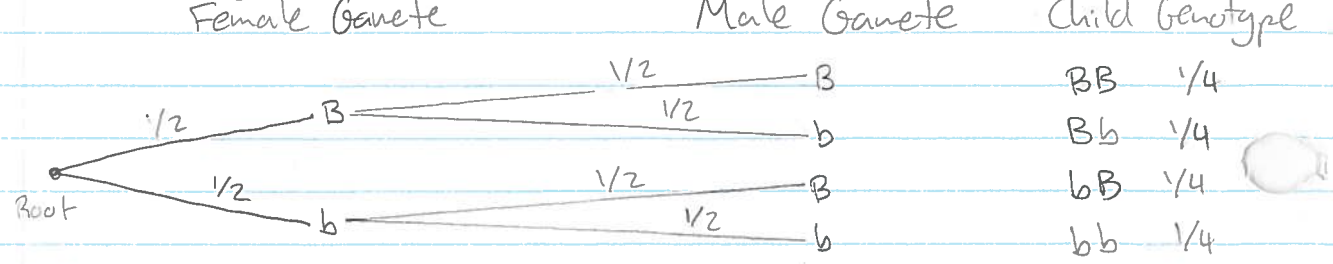
3) The classical method
- this is used when we have equally likely events
- $P(A) = \frac{n(A)}{n(S)}$
- ex: 52 cards, $n(S) = 52$, pick randomly one card
  $A = \{$pick an ace$\}$, $n(A) = 4$

$P(A) = \frac{n(A)}{n(S)} = \frac{4}{52} = \frac{1}{13}$

Section 2.1: Tree Diagrams

A <u>tree diagram</u> is a graphical method used to represent the outcomes of a random experiment
- start with a common "root"
- draw a branch for each possible outcome

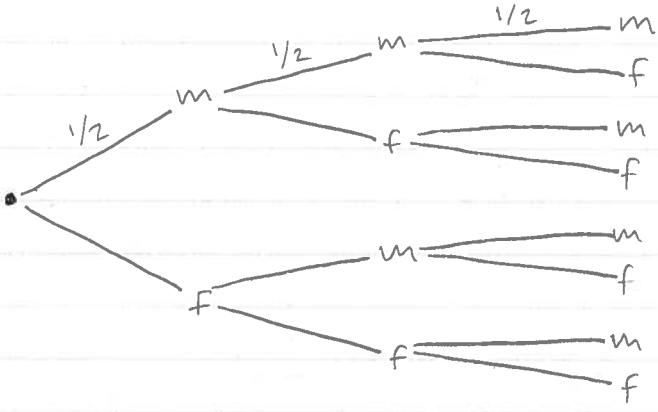
<u>Ex 1</u> each member of a couple has alleles for both brown (B) and blue (b) eyes. Their genotype is Bb, therefore they are heterogenous. B is dominant and b is recessive. The four possible child genotypes are: BB, Bb, bB, bb. Bb and bB are the same result occuring from two different processes.
Tree diagram for the child's genotype:



Probability that the child has blue eyes

$P(bb) = \frac{1}{4} = 0.25 = 25\%$

Ex 2 3 children, male (m), female (f). What is the birth order



Birth Orders	Probabilities
mmm	1/8
mmf	1/8
mfm	1/8
mff	1/8
fmm	1/8
fmf	1/8
ffm	1/8
fff	1/8

a) find the probability that the couple has only boys.
solution: 1/8

b) find the probability that there is at least one boy.
solution: P(at least one) =
P({exactly one boy})
+ P({exactly two boys})
+ P({exactly three boys})
= 7/8

method 2: $1 - P(\text{only girls}) = 1 - 1/8 = 7/8$

MAT2379A 09/09/2013

Practice Problems
1) problems at the end of each chapter
2) Extra problems in chapter 8 (probability) and chapter 16 (statistics)
3) course website, link to self-test problems

Assignment 1 (problems from book) due sept. 20th

2.1 Tree Diagrams

Ex 3: The alleles for normal skin pigmentation S is dominant over albinism (s). The allele for free (unattached) earlobes F is dominant over attached earlobes f. A mother has normal skin pigmentation and attached earlobes. Her possible genotypes are SS and Ss for skin pigmentation and ff for earlobes A father is albino and has free earlobes. His possible genotypes are ssFf or ssFF. There are four possible mating cases:

Case 1: woman is SSff man is ssFF
Case 2: woman is Ssff, man is ssFf
Case 3: woman is Ssff, man is ssFF
Case 4: woman is Ssff, man is ssFf

Consider case 3: tree diagram which gives all possible genotypes of their child

Female (skin)	Male (skin)	Female (earlobes)	Male (earlobes)	Child Genotype
S	s	f	F	SsfF
s	s	f	F	ssfF

P(child has normal skin pigmentation and free earlobes)
= P(SsfF)
= 1/2 = 0.5 = 50%

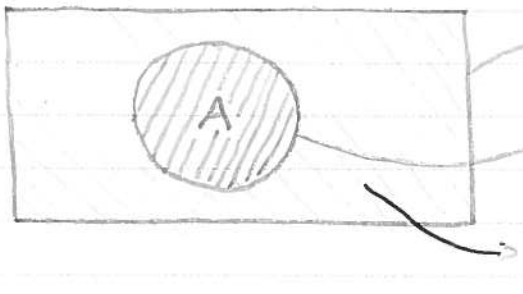
Exercise: Do same for case 1, 2, and 4

Ch. 3 - Axioms of Probability

3.1 - Venn Diagrams

- a Venn Diagram is a graphical method used for representing sets

- S = sample space (set of all possible outcomes of a random experiment); represented by a rectangle



An event A is represented as a closed curve inside the rectangle.

- the compliment of A is the event "A fails" and is denoted by A'

$$P(A) + P(A') = 1 = P(S)$$

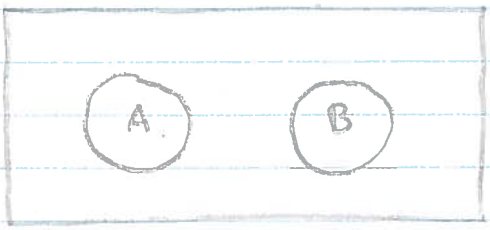
Ex 1: A family has 3 children.
A = {family has only boys}
A' = {family has at least one girl}
P(A) = P{mmm} = 1/8 (see diagram from previous example)
P(A') = P{fmm, mfm, mmf, ffm, fmf, mff, fff} = 7/8
= 1 - P(A) = 1 - P{mmm} = 1 - 1/8 = 7/8

Note = at times it may be easier to compute $P(A')$ than $P(A)$ itself. If that's the case, the use:

$$P(A) = 1 - P(A')$$

Let's look at 2 events A and B. We have two cases:

1) If the event A and B cannot occur at the same time then the curves representing A and B do NOT overlap



"A or B occurs" is denoted by $A \cup B$ (union)

$$P(A \cup B) = P(A) + P(B)$$

Ex2 - In Canada we have the following blood distributions: 42% type A blood
9% type B blood
3% type AB blood
46% type O blood

A new patient is admitted into a hospital and needs a blood transfusion. What is P(patient is type A or type B blood)?

A = event that the person has type A blood
B = event that the person has type B blood

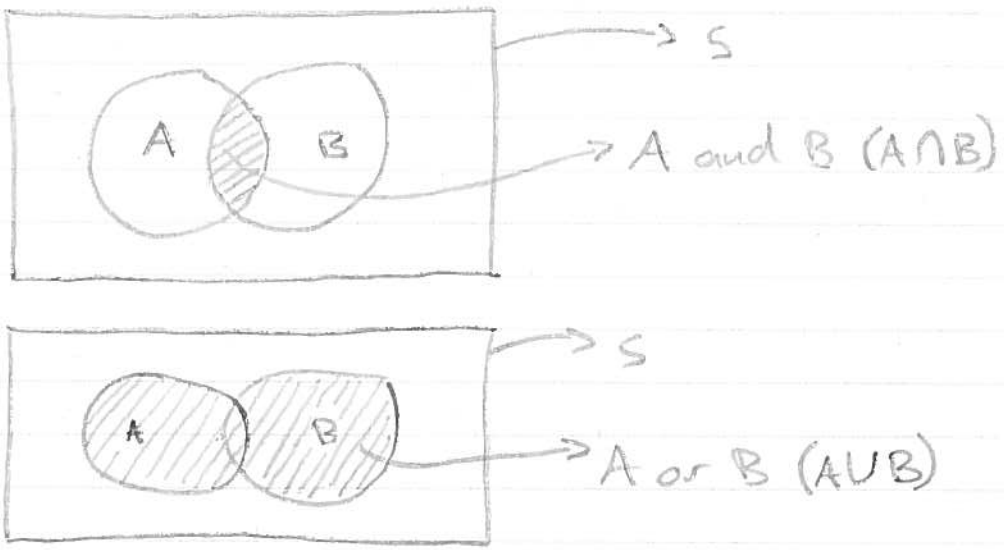
A and B are mutually exclusive: they do not occur at the same time

$P(A \text{ or } B) = P(A \cup B)$
$= P(A) + P(B)$
$= 0.42 + 0.09$
$= \underline{\underline{0.51}}$

intersect
The event $A \cap B$ corresponds to "A and B occur".
In the prev. example,
$P(A \cap B) = P(A \text{ and } B) = \text{impossible event} = \emptyset$

Notice: if A and B are mutually exclusive, then
$P(A \cap B) = 0$

2) Consider two events A and B which may occur at the same time. In this case, the 2 curves representing A and B are overlapping

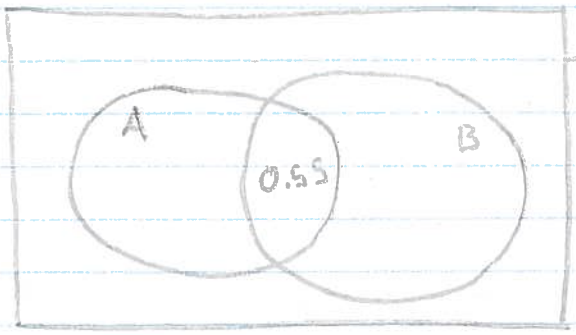


Ex 3
100 german men
→ 70 have blond hair (event A)
→ 77 have blue eyes (event B)

A and B are not mutually exclusive, ie. some people in the group have both characteristics

If we select randomly a person in this group and we let A the event that the person has blond hair, then $P(A) = 70/100 = 0.7$, and we let B the event that the person has blue eyes, then $B = 77/100 = 0.77$.

Suppose 55 people have both blond hair and blue eyes. Then $P(A \cap B) = 55/100 = 0.55$



$P(A) = 0.7$
$P(B) = 0.77$
$P(A \cap B) = 0.55$

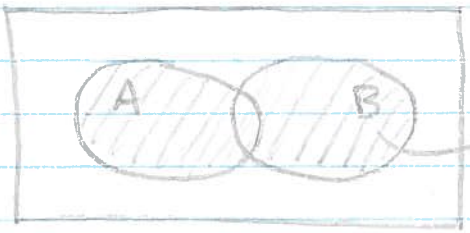
32-Addition Rule

if $A \cap B = \emptyset$, then we saw that $P(A \cup B) = P(A) + P(B)$

if $A \cap B \neq \emptyset$, then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

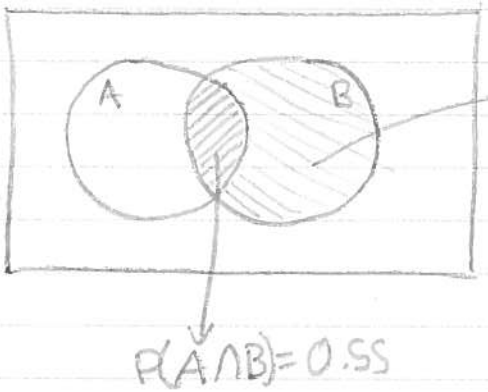
Ex 3 - What is the probability that a randomly selected person in this group has blond hair or blue eyes?

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
$= 0.70 + 0.77 - 0.55$
$= 0.92$



→ $A \cup B$ has a prob. of 0.92

b) What is the probability that a randomly selected person does not have blond hair but does have blue eyes.

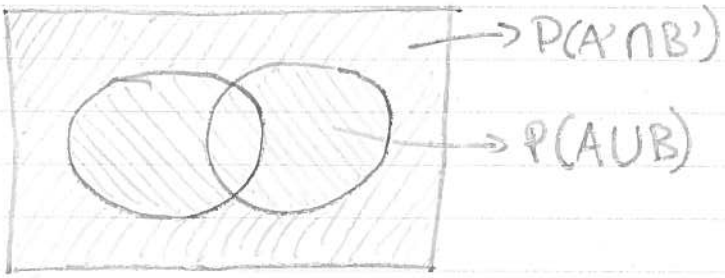


$P(B) = 0.77$
$P(A \cap B) = 0.55$

$P(B \cap A') = 0.77 - 0.55 = 0.22$
$P(A \cap B') = 0.70 - 0.55 = 0.15$

In math terms, we write $P(A \cap B') = P(A) - P(A \cap B)$

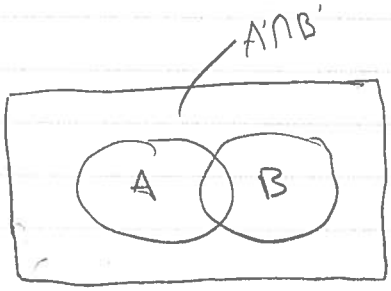
d) Probability that someone in this group does not have blond hair and blue eyes:
$P(A' \cap B') = 1 - P(A \cup B) = 1 - 0.92 = 0.08$



MAT2379A 11/09/2013

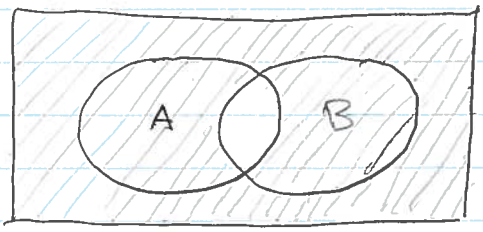
Ex from last class
- 100 ppl

$A = \{\text{blond}\}$ $P(A) = 0.70$
$B = \{\text{blue}\}$ $P(B) = 0.77$
$A \cap B = \{\text{blond and blue}\}$ $P(A \cap B) = 0.55$



$P(\underbrace{\text{not blond}}_{A'}, \underbrace{\text{not blue}}_{B'}) = P(A' \cap B')$
$= 1 - P(A \cup B)$

$P(\text{Not blond OR NOT blue}) = P(A' \cup B') = 1 - P(A \cap B)$



De Morgen Laws

$(A \cap B)' = A' \cup B'$
$(A \cup B)' = A' \cap B'$

Ch. 4 - Conditional Probability

- Given that (a person is blond)$_A$, what is the prob. that this (person has blue eyes.)$_B$

$P(B \mid \text{given that } A \text{ already happened})$

↳ syntax: bar |

$= P(B|A) = 55/70 = 0.79$ note $P(A) = 0.70$, $P(A \cap B) = 0.55$

$P(B|A)$ is called the conditional probability of B given A:

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

Ex 1

43% smoke (A)
70% stressed (B)
85% smoke or are stressed (A∪B)

a) What is the probability that a randomly chosen person smokes and is stressed?

$P(A \cap B) = ?$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
$0.85 = 0.43 + 0.70 - P(A \cap B)$
$P(A \cap B) = 0.43 + 0.70 - 0.85$
$= 0.28$

b) What is the probability that a person is stressed given that this person smokes?

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.28}{0.43} = 0.651$$

c) Reading exercise: what is the probability that a person smokes given that this person is not stressed?

Diagnostic Tests

- a diagnostic test is a medical test given to a patient to detect the presence of a disease, such as for HIV, DOWN syndrom, prostate cancer, etc...

- there are four possible situations:
  1) the subject is a TRUE positive and the test result is also positive = no error
  2) the subject is a TRUE negative and the test result is positive = error

~~3) The subject is a TRUE positive~~
This conditional probability is called the
false-positive rate of the test

P(test is positive | patient is true negative) = α
-you want ~~this~~ this to be small.

3) The subject is a TRUE negative and the test result is negative = no error

4) The subject is a TRUE positive and the test result is negative = error

P(test is negative | patient is true positive) = β
-you want this to be very, very small
-β is called false-negative rate of the test

	True + (U+)	True - (U-)
Test + (T+)	correct	Error
Test - (T-)	Error	correct

Specificity = P(Test is -ve | Subject is -ve)
Sensitivity = P(Test is +ve | Subject is +ve)

Positive predictive value (PPV)
= P(subject +ve | test +ve)

negative predictive value (NPV)
= P(subject -ve | test -ve)

These four values should be high; >90 % for a good test.

Ex - DOWN syndrome test

300 pregnant women decided to take the test
5 obtained a +ve test result
295 obtained a -ve test result

	True + Down Syn	True - Healthy Baby	Total
test +	4	1	5
Test -	3	292	295
Total	7	293	300

From the 5 women who obtained a +ve test result:
- 4 were truly carrying a baby with down syndrome
- 1 was carrying a healthy baby.

From the 295
- 3 were carry a child with Down syndrome.
- 292 healthy

$$\alpha = P(\text{test}+ \mid \text{true}-) = \frac{P(\text{test}+ \cap \text{true}-)}{P(\text{true}-)} = \frac{1/300}{293/300} = \frac{1}{293} \approx 0.003$$

$$\beta = P(\text{test}- \mid \text{true}+) = \frac{P(\text{test}- \cap \text{true}+)}{P(\text{true}+)} = \frac{3/300}{7/300} = \frac{3}{7} = 0.429$$

$$PPV = P(\text{true}+ \mid \text{test}+) = \frac{P(\text{true}+ \cap \text{test}+)}{P(\text{test}+)} = \frac{4/300}{5/300} = \frac{4}{5} = 0.80$$

4.2 - Multiplication Rule and Total Probability

$P(B|A) = \frac{P(A \cap B)}{P(A)}$, we rewrite this as:

$P(A \cap B) = P(A) \cdot P(B|A)$ → multiplication rule

ex
2% of population has diabetes (A), $P(A) = 0.02$
(of these,
50% are unaware of their condition (B)

$P(B|A) = 0.50$

What is the probability that a randomly selected person has diabities and is unaware of the condition.

$$P(A \cap B) = P(A) \cdot P(B|A)$$
$$= 0.02 \cdot 0.50$$
$$= 0.01$$

∴ 1% of the population has diabetes and is unaware of it.

ex In a population of people of age 50+, 56% are women. Of these, 12% have arthritis. In the pop. of men 50+, 10% have arthritis. What is the probability that a person in this population has arthritis?

A = "person is a woman", $P(A) = 0.56$
A' = "person is a man", $P(A') = 0.44$
B = "person has arthritis"
$P(B|A) = 0.12$
$P(B|A') = 0.10$

$$P(B) = P(\text{arthritis})$$
$$= P(\text{arthritis and woman}) + P(\text{arthritis and man})$$
$$= P(B \cap A) + P(B \cap A')$$
$$= P(A) \cdot P(B|A) + P(A')\, P(B|A')$$
$$= 0.56 \times 0.12 + 0.44 \times 0.10$$
$$= 0.0672 + 0.044$$
$$= 0.1112$$

MAT2379-A 16/09/2013

Problem 4.4 c,d

Prevelence of a disease = percentage of diseased = P(True +)
See example 4.10 (p.40), forget abut the expanded form.

Problem 4.1

Hint: consider seperatly the male and female populations.

$P(\text{lung cancer} \mid \text{smoker}) = 0.172$
$P(\text{lung cancer} \mid \text{non-smoker}) = 0.013$ } male population
$P(\text{cancer}) = 0.522$

Total Probability Rule

$$P(A|B)P(B) + P(A|B')P(B') = P(A)$$

4.3 Bayes Rule

Suppose that we know $P(B)$ (hence $P(B')$)
$P(A|B)$
$P(A|B')$

We can compute:

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

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B')P(B')}$$ → Bayes Rule

- used the multiplaction rule in the nominator
- used the total probability rule in the denominator

Example 5

A screening test is conducted to detect the presence of a medical condition. If someone has a disease, the probability that the test is +ve is 80%, ie.

$$P(\text{Test} + | \text{True} +) = 0.80$$

If someone does not have the disease, the probability that the test is -ve is 88%, ie.

$$P(\text{Test} - | \text{True} -) = 0.88$$

5% of the population has the disease (prevelence), ie.

$$P(\text{True} +) = 0.05$$

a) What is the prob. that a randomly chosen person has a +ve test?

$P(\text{Test}+) = ?$

Use total prop. rule:

$P(\text{Test} + | \text{True} +) = 0.80$
$P(\text{Test} + | \text{True} -) = 1 - P(\text{Test} - | \text{True} -) = 1 - 0.88 = 0.12$
$P(\text{True} +) = 0.05$
$P(\text{True} -) = 1 - P(\text{True} +) = 1 - 0.05 = 0.95$

$$P(\text{Test}+) = P(\text{Test}+|\text{True}+)P(\text{True}+) + P(\text{Test}+|\text{True}-)P(\text{True}-)$$
$$= (0.80)(0.05) + (0.12)(0.95)$$
$$= 0.154$$

b) What is the probability that a randomly chosen person has the disease given that this person had a +ve test result?

$$P(\text{True}+|\text{Test}+) = \frac{P(\text{True}+ \wedge \text{Test}+)}{P(\text{Test}+)} \longrightarrow \text{PPV}$$
$$= \frac{P(\text{Test}+|\text{True}+)P(\text{True}+)}{P(\text{Test}+)}$$
$$= \frac{(0.80)(0.05)}{(0.154)}$$
$$= 0.26$$

Note that the probability
$$P(\text{Test}+|\text{True}+) \neq P(\text{True}+|\text{Test}+)$$
$$0.80 \neq 0.26$$

80% is the percentage of ppl who will obtain a +ve test result, in the group of diseased ppl.

26% is the percentage of ppl who are diseased, in the group of ppl who obtained a +ve test result.

Ch. 5 - Independence

- Recall that events A and B are <u>mutually exclusive</u> if they can not happen at the same time.

- We say that events A and B are <u>independent</u> if the fact that one of them (say A) happened does not influence the prob. that the other event (B) will happen. ie. $P(B|A) = P(B)$ (1)

or equivalently as:

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

which in turn is equivalent to:

$$P(B \cap A) = P(A)\,P(B) \quad (2)$$

To show that events A and B are <u>independent</u> we have to show that (1) or (2) holds.

<u>Example 1</u>

350 women aged 50+
214 do not exercise; (A); $P(A) = 214/350$
75 have osteoporosis; (B); $P(B) = 75/350$
49 do not exercise and have osteoporosis; $(A \cap B)$; $P(A \cap B) = 49/350$

Is the fact that a woman has signs of osteoporosis <u>independent</u> of the fact that she does not exercise.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{49/350}{214/350} = \frac{49}{214} = 0.229$$

$$P(B) = 75/300 = 0.214$$

$$P(B|A) \neq P(B)$$
$$0.229 \neq 0.214$$

$\therefore P(B|A) \neq P(B)$ $\therefore$ the fact that a woman has signs of osteoporosis <u>is not independent</u> of the fact that she does not exercise.

Interpretation: the fact that $P(B|A) > P(B)$ means that the fact that a woman does not exercise <u>increases</u> her chances of developing osteoporosis.

Example 2

50% of population is male; (A), $P(A) = 0.50$
68% of population drinks to some extent; (B), $P(B) = 0.68$
38.5% of population drinks and male; $(A \cap B)$, $P(A \cap B) = 0.385$

Is the person drinking status <u>independent</u> of gender? Is A independent of B? See if (2) holds.

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

Left side	Right sides
$= P(A \cap B)$	$= P(A)\,P(B)$
$= 0.385$	$= (0.50)(0.68)$
	$= 0.34$

$\therefore$ LS $\neq$ RS $\therefore$ A and B are not independent; B is dependent of A

Remark: $P(\text{drinks} \mid \text{male}) = P(B|A) = \dfrac{P(B \cap A)}{P(A)} = \dfrac{0.385}{0.50} = 0.77$
$P(B|A) > P(B)$, so the fact that someone is male increases the chances that they drink.

Example 3

Three diagnostic tests are run on the same persons: T_1, T_2, T_3.
The results of the tests are Independent of each other.
$P(T_1 \text{ is correct}) = 0.90$; (A)
$P(T_2 \text{ is correct}) = 0.85$; (B)
$P(T_3 \text{ is correct}) = 0.80$; (C)

a) what is the prob. that all three test will give a correct result?

Note: We say that A, B, C are independent if A, B are indep. A, C are indep. B, C are ind. and $P(A \cap B \cap C) = P(A)\,P(B)\,P(C)$

$$P(A \cap B \cap C) = P(A)\,P(B)\,P(C)$$
$$= (0.90)(0.85)(0.80)$$
$$= 0.612$$

b) What is the prob. that at least one test will result in error?
This is the compliment of the event D: "all tests give correct results". $P(D) = 0.612$

$$P(D') = 1 - P(D) = 1 - 0.612 = 0.388$$

MAT2379 10/09/2013

Download formula sheet from website.

Ch. 5 - Independence Cont.

We say that events A and B are independent if

$$P(B|A) = P(B) \quad [1]$$

or equivalently

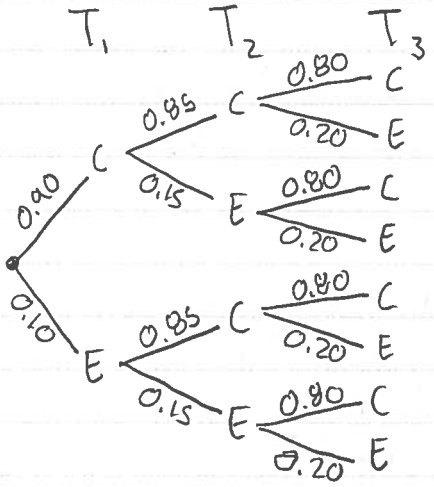
$$P(A \cap B) = P(A)P(B) \quad [2]$$

Ex 3) Tests T_1, T_2, T_3 are independent

$$\begin{cases} P(T_1 \text{ correct}) = 0.90 \\ P(T_2 \text{ correct}) = 0.85 \\ P(T_3 \text{ correct}) = 0.80 \end{cases}$$

E = error
C = correct

c) What is the probability that exactly two tests will be correct.



Outcomes	Probability
CCC	0.612
CCE	0.153
CEC	0.108
CEE	0.027
ECC	0.068
ECE	0.017
EEC	0.012
EEE	0.003

We are interested in event A that exactly two tests have a correct result

$$P(A) = P(T_1 C, T_2 C, T_3 E) + P(T_1 C, T_2 E, T_3 C) + P(T_1 C, T_2 C, T_3 C)$$
$$= 0.153 + 0.108 + 0.068$$
$$= 0.329$$

Ch 6 - Discrete Random Variables

6.1 - Definition

- a measurment made on a subject selected randomly from a large population is called a <u>random variable</u>
- such an object is denoted with a capital letter, such as: X, Y, Z, etc...
- for each random variable X, we are interested in propabilities associated to its value, ex:

  $\underbrace{P(X \leq 180)}_{\text{event A}}$; X = blood pressure

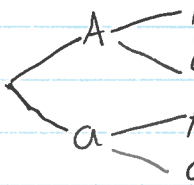
- the values which are assumed by X are denoted by lowercase letters, x, y, z, etc... (numerical values)
- random variables can be of 2 types:
  + discrete (Finit set of possible values, listable)
  + continuous (infinitly many possibilities, unlistable)
- Discrete random variables are those which take only finitely many values.

ex) - sex of a newborn infant (X); the possible values are male or female

ex) - the number of worker bees in a honeybee society

ex) - blood type (X); possible values: A, B, AB, O

ex) - the number of genes A in an offspring of two heterozygous individuals Aa,



AA	1/4	Possible genotypes:
Aa	1/4	P(Aa) = 1/2
aA	1/4	P(AA) = 1/4
aa	1/4	P(aa) = 1/4

X can take the values 0, 1, 2:

$P(X=0) = P(aa) = 1/4$

$P(X=1) = P(Aa) = 1/2$

$P(X=2) = P(AA) = 1/4$

x	0	1	2
P(X=x)	1/4	1/2	1/4

Examples of continuous random values; they take infinitely many values, usually in a range =
1) weight or height of a patient
2) blood pressure
3) weight gain/loss

The function $f(x) = P(X = x)$ is called the probability mass function or probability density function of the discrete random variable X

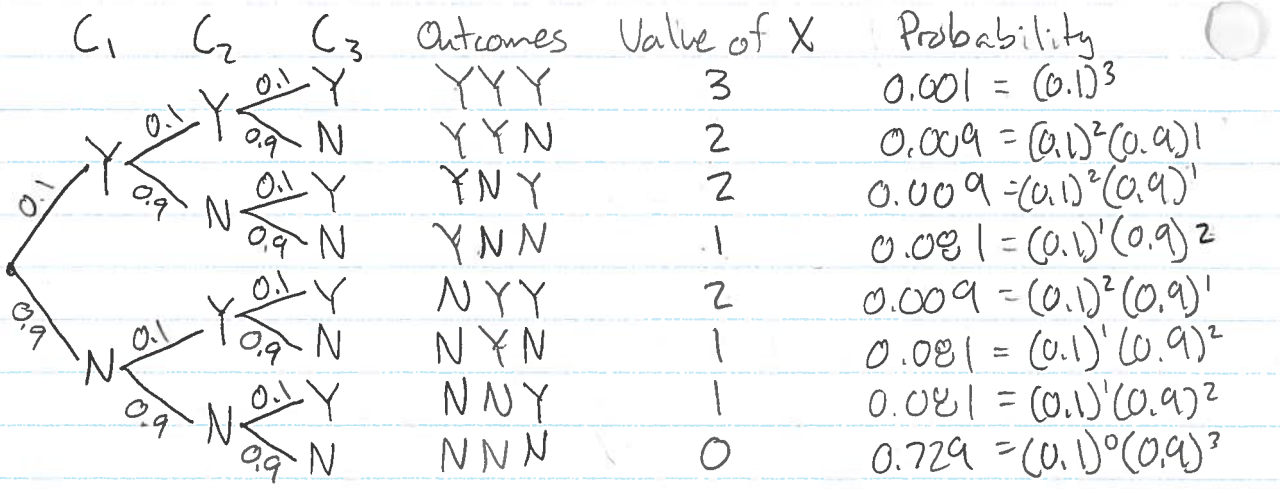
Note :- if x is an impossible value of X, then
$P(X = x) = 0$
(impossible event)
- $\sum P(X = x) = 1$ → all possible x
- $P(X = x)$ is in $[0, 1]$

Ex 1) In a certain population, one child in ten has a high blood lead level (ie. the prob. that a child has this condition is 0.1). Consider a randomly chosen group of 3 children in this population.

Let X be the number of children in the group of three who have a high blood lead level. X is a discrete random variable with possible values: 0, 1, 2, 3

We would like to compute the probabilities
$P(X = x)$ where $x = 0, 1, 2$, or 3

We use a tree diagram; C_1 = child 1, C_2 = child 2, C_3 = child 3
Y = yes, high blood lead level, N = no, normal blood lead level.



Summary!

$f(0) = P(X=0) = (0.1)^0(0.9)^3 = 0.729$

$f(1) = P(X=1) = 3(0.1)(0.9)^2 = 0.243$

$f(2) = P(X=2) = 3(0.1)^2(0.9) = 0.027$

$f(3) = P(X=1) = (0.1)^3 = 0.001$

★ 3 = # of ways of choosing one child in a sample of 3

★ 3 = # of ways of choosing two children in a sample of 3

x	0	1	2	3
$P(X=x)$	0.729	0.243	0.027	0.001

Note that $\sum_{x=0}^{3} f(x) = 1$

a) What is the probability that at least two children in the group have high blood lead level?

$$P(X \geq 2) = P(X=2) + P(X=3)$$
$$= (0.027) + (0.001)$$
$$= 0.028$$

Warning! Make sure you understand that this is not the same as $P(X>2) = P(X=3)$

b) What is the probability that at most two children have high blood lead levels?

$$P(X \leq 2) = P(X=0) + P(X=1) + P(X=2)$$
$$= (0.729) + (0.0243) + (0.027)$$
$$= 0.999$$

note compare with $P(X < 2)$

MAT 2379A 23/09/2013

6.1 - Definition of Discrete Random Variables

X = random quantity or measurement; it takes only countably many values.

We are interested in $P(X = x)$, x: numeric value

Ex 2

X = number of persons per day who are seeking emergency room treatment unnecessarily in a small hospital

value	0	1	2	3	4	5
frequency	297	25	18	14	8	3

total = 365 days

x	0	1	2	3	4	5
P(X=x)	0.81	0.07	0.05	0.04	0.02	0.01

→ sum = 1

What is the probability that in a randomly chosen day there will be at least 4 persons seeking emergency room treatment unnecessarily?

$P(x \geq 4)$
$= P(x=4) + P(x=5)$
$= 0.02 + 0.01$
$= 0.03$

Expectation

The expectation (or expected value) of a discrete random variable X is given by:

$$E(x) = \sum_{all\,x} x f(x) = \sum_{all\,x} x P(X=x)$$

$E(x)$ is interpreted as the average value of X

Ex 1 (cont.)

x	0	1	2	3
$P(X=x)$	0.729	0.243	0.027	0.001

Compute the average number of children with a high blood lead level, in a sample of three

$E(x) = \sum x f(x)$
$= (0)(0.729) + (1)(0.243) + (2)(0.027) + (3)(0.001)$
$= 0 + 0.243 + 0.054 + 0.003$
$= 0.300$

note = it's ok for an average to have a decimal form even though the initial values are integers.

Variance

The variance of a discrete random variable X is a measure of the amount of dispersion of X around its average. Is given by:

$$\text{Var}(X) = \sum_{\text{all } x} (x - E(x))^2 P(X = x)$$

Measurement unit for Var(x) is: (original unit)^2
The root of the variance is called the standard deviation

$$\sqrt{\text{Var}(x)}$$

Notation with greek letters:

$E(x) = \mu$, pronounced [mu:]

$\sqrt{\text{Var}(x)} = \sigma$, pronounced [sigma]

Therefor $\text{Var}(x) = \sigma^2 \geq 0$
Note μ can be +ve or -ve but σ is always +ve

Ex 1 (cont.)

x	0	1	2	3
$P(X=x)$	0.729	0.243	0.027	0.001

Compute the variance and the standard deviation for the number X of children with high blood lead level, in a sample of 3.

$$\text{Var}(x) = (0-0.3)^2(0.729) + (1-0.03)^2(0.243) + (2-0.3)^2(0.027) + (3-0.3)^2(0.001)$$
$$= 0.27 \text{ children}^2$$

Standard deviation of X is

$\sqrt{\text{Var}(x)} = \sqrt{0.27} = 0.5196$ children

Shortcut for variance:

$$\text{Var}(x) = \sum x^2 P(X=x) - [E(x)]^2$$

Ex 1 (cont.)

$$\text{Var}(x) = (0)^2(0.729) + (1)^2(0.243) + (2)^2(0.027) + (3)^2(0.001) - (0.3)^2$$

$$\sigma^2 = 0.36 - 0.09$$

$$= 0.27 \text{ children}^2$$

Cumulative Distribution Function (cdf)

$$F(x) = P(X \leq x) = \sum_{\text{all } y \leq x} P(Y = y)$$

Ex 1 (cont.)

Compute the cdf of X

$F(0) = P(X \leq 0) = P(X=0) = 0.729$

$F(1) = P(X \leq 1) = P(X=0) + P(X=1) = 0.729 + 0.243 = 0.972$

$F(2) = P(X \leq 2) = P(X=0) + P(X=1) + P(X=2)$

$= 0.729 + 0.243 + 0.027 = 0.999$

$F(3) = P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3)$

$= 0.729 + 0.243 + 0.027 + 0.001 = 1$

x	0	1	2	3
$F(x)$	0.729	0.972	0.999	1

Note: we can recover the original table of probabilities $f(x) = P(X=x)$ from the table of $F(x) = P(X \leq x)$

If x is an impossible value, then $P(X=x) = 0 = f(x)$

Ex: $x = 2.5$, $P(X = 2.5) = 0$

But even if x is an impossible value of X, $F(x)$ may not be 0!

Ex: $x = 2.5$

$$\begin{aligned} F(2.5) &= P(X \leq 2.5) \\ &= P(X=0) + P(X=1) + P(X=2) \\ &= 0.999 \end{aligned}$$

6.2 Binomial Distribution

- aka binomial random variable
- this is an important example of a discrete random variable.
- we encounter it in the following situation:
  1) we have an experiment which consists of n identical and independent trials. (ex = flips of coins)
  2) each trial has a result which can be classified as a success or failure. (ex = heads → success; tails → failure)
  3) for each trial, the probability of success is the same value p
  4) we are interested in the total number X of "successes" (obviously, X is a discrete random variable which takes values in the set $0, 1, 2, \ldots, n$)

By definition we say that X is a binomial random variable with n trials and a probability p of success.

Ex 1

each child (in a sample of 3) is a "trial" "success" means that the child has high blood lead level, and the probability p of success was 0.1

Therefore, X is a binomial $(n=3, p=0.1)$

$$\begin{cases} P(X=0) = \binom{3}{0}(0.1)^0(0.9)^3 \\ P(X=1) = \binom{3}{1}(0.1)^1(0.9)^{3-1} \\ P(X=2) = \binom{3}{2}(0.1)^2(0.9)^{3-2} \\ P(X=2) = (0.1)^3(0.9)^{3-3} \end{cases}$$

$\binom{3}{0}$: number of ways of choosing 1 child among 3.

In general, for any binomial random variable X with n trials and probability p of success,

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

$\binom{n}{x}$ → number of ways of choosing x subjects among n subjects, called the binomial coefficient

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} = \text{binomial coefficient (read } n \text{ choose } x)$$

values $\binom{n}{x}$ are given by Table 17.1

MAT2379 25/09/2013

Assignment 2 due Oct. 4th

6.2 - Binomial Random Variables

X = number of "success" in a sequence of n trials, each trial has probability p. of "success"

$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$, $\binom{n}{x}$ is defined in §2.2

$\binom{n}{x} = \frac{n!}{x!(n-x)!}$ → given in table 17.1 $\binom{n}{x} = \binom{n}{n-x}$

ex 3

Probability of germination of a seed is $p=0.8$. We plant $n=3$ seeds and we let X = total number of germinated seeds. X is a binomial random variable with $\begin{cases} n=3 \\ p=0.8 \end{cases}$

Calculate $P(X=x)$ for $x=0,1,2,3$

$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{3}{x} 0.8^x (0.2)^{3-x}$

$P(X=0) = \binom{3}{0} 0.8^0 (0.2)^{3-0}$
$= (1)(1)(0.008)$
$= 0.008$

$P(X=1) = \binom{3}{1} 0.8^{(1)} (0.2)^{3-(1)}$
$= (3)(0.8)(0.04)$
$= 0.96$

$P(X=2) = \binom{3}{2} 0.8^2 (0.2)^{3-2}$
$= (3)(0.64)(0.2)$
$= 0.384$

$\frac{3!}{1!(3-1)!} = \frac{6}{1(2)} = 3$

$\frac{3!}{2!(3-2)!} = \frac{6}{2(1)} = 3$

Ex 4

85% of the population has RH positive blood. We have a sample of 6 persons. X = number of persons (in the sample of six) with RH positive blood.

X has a binomial distribution (it is a binomial random variable), with $\begin{cases} n=6 \\ p=0.85 \end{cases}$

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{6}{x} 0.85^x 0.15^{6-x} \quad \text{for } x = 1, 2, \ldots, 6$$

$$\begin{aligned} P(X=2) &= \binom{6}{2} 0.85^2 0.15^{6-2} \\ &= (15)(0.7225)(0.15^4) \\ &= 0.0055 \end{aligned}$$

The expected value of X (binomial (n,p)) is:

$$E(X) = np$$

The variance of X is:

$$Var(X) = np(1-p)$$

Ex 4 cont.

$$E(X) = np = (6)(0.85) = 5.1$$
$$Var(X) = np(1-p) = (6)(0.85)(1-(0.85)) = 0.765$$

Example of a calculation with a Discrete random variable.

$$P(5 \leq X \leq 10) =$$
$$= P(X=5) + P(X=6) + P(X=7) + P(X=8) + P(X=9) + P(X=10)$$

Suppose that the probability $P(X=x)$ are not given, but we are given $P(X \leq x)$ for all x

$$P(5 \leq X \leq 10) = P(X \leq 10) - P(X \leq 4)$$

Practise

Compute $P(3 < X < 7) = P(X \leq 6) - P(X \leq 3)$

$P(6 \leq X < 11) =$

$P(8 < X \leq 12) =$

Ch. 7) Continuous Random Variables

7.1 - Definitions

A random variable is <u>continuous</u> if it takes infinitely many values, usually in a range. We have:

$P(X=x) = 0$ for all → difficulty

ex) weight of a person, blood pressure, weight gain/loss.

Even though we can not work with $P(X=x)$ (they are 0), we can still work with X by focussing on:

$$P(a \leq X \leq b) = \int_a^b f(x)dx \qquad (1)$$

This function $f(x)$ plays the same role as the function:

$f(x) = P(X = x)$ in the discrete case.

From (1), if we take $b = a + \varepsilon$ (ε small)

$$\frac{1}{\varepsilon} P(a \le X \le a+\varepsilon) = \frac{1}{\varepsilon} \int_a^{a+\varepsilon} f(x)\,dx$$

$$\xrightarrow[\varepsilon \to 0]{} fa$$

Hence $\boxed{P(a \le X \le a+\varepsilon) \approx \varepsilon f(a)}$ → density function

properties: $f(x) \ge 0$, $\int_{-\infty}^{\infty} f(x)\,dx = 1$

Expect value of X is:

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

Variance:

$$\sigma^2 = \text{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$$

Cumulative distribution function of X:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(y)\,dy$$

MAT 2137 30/09/2013

Hint for Problem 6.8

- Y = number of litters (in a sample of 10) which have at least 2 cubs

- Y counts the number of "successes"
- "success" = "at least 2 cubs"
- Y has a binomial distribution
  - $n = 10$
  - p = Prob(at least 2 cubs)
    $= P(X \geq 2)$ where X is from problem 6.7

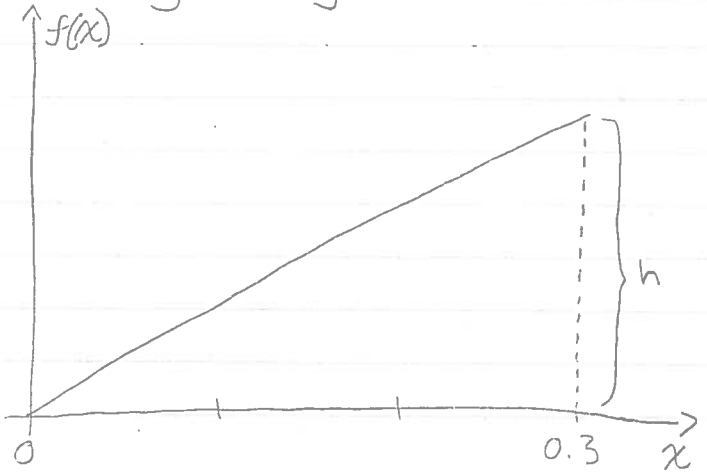
Assignment 3

- due friday, Oct. 11th at 3:00PM
- based on ch 7 and one question ch 6

7.1 - Definition

Continuous random variable

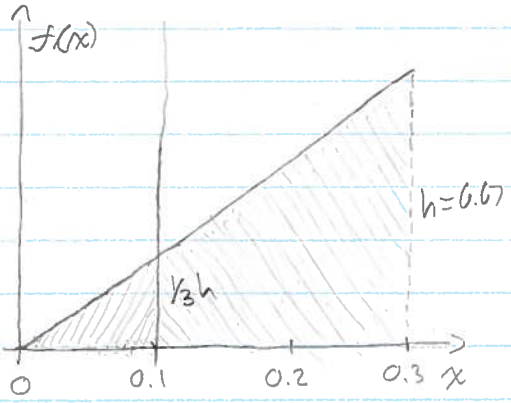
ex) X = quantity of cubic centimeters (cc's) of a drug prescribed for the control of epileptic seizures.
X takes values in the range [0, 0.3], with the following density function:



a) Find "h".
- area under the graph of f(x) should be 1

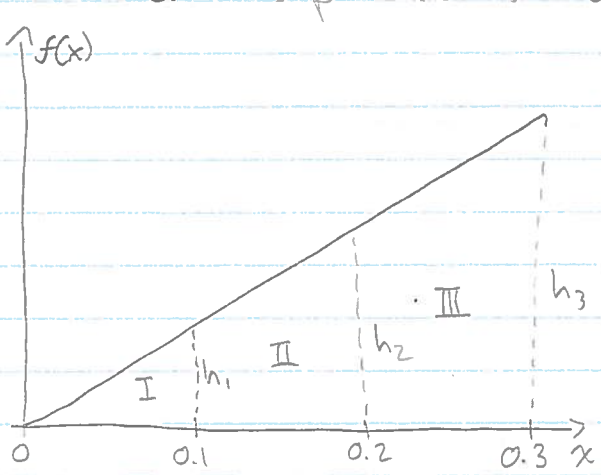
$A = (w h)/2$
$(1) = (0.3)(h)/2$
$h = 6.67$

b) Shade and find the area corresponding to the probability that x is at least 0.1.



$$P(x \geq 0.1) = 1 - P(x < 0.1)$$
$$= 1 - \frac{(0.1)(\frac{1}{3}h)}{2}$$
$$= 1 - \frac{(0.1)(\frac{1}{3})(6.67)}{2}$$
$$= 0.89$$

c) What are the probabilities corresponding to areas I, II, III?



$h_3 = 6.67$
$h_2 = (\frac{2}{3})(6.67) = 4.44$
$h_1 = (\frac{1}{3})(6.67) = 2.22$

for area I:
$P(x \leq 0.1) = 0.11$

for area II:
$P(0.1 \leq x \leq 2)$
$= P(x \leq 2) - P(x \leq 0.1)$
$= \frac{(0.2)(4.44)}{2} - (0.11)$
$= 0.44$

for area III:
$P(0.2 \leq x \leq 0.3)$
$= 1 - P(x \leq 0.2)$
$= 1 - (0.44)$
$= 0.56$

d) What is the probability that exactly 0.2 cc's are prescribed?
$P(x = 0.2) = 0$

Summary: for a continuous variable X

→ $P(X = x) = 0$ for any x

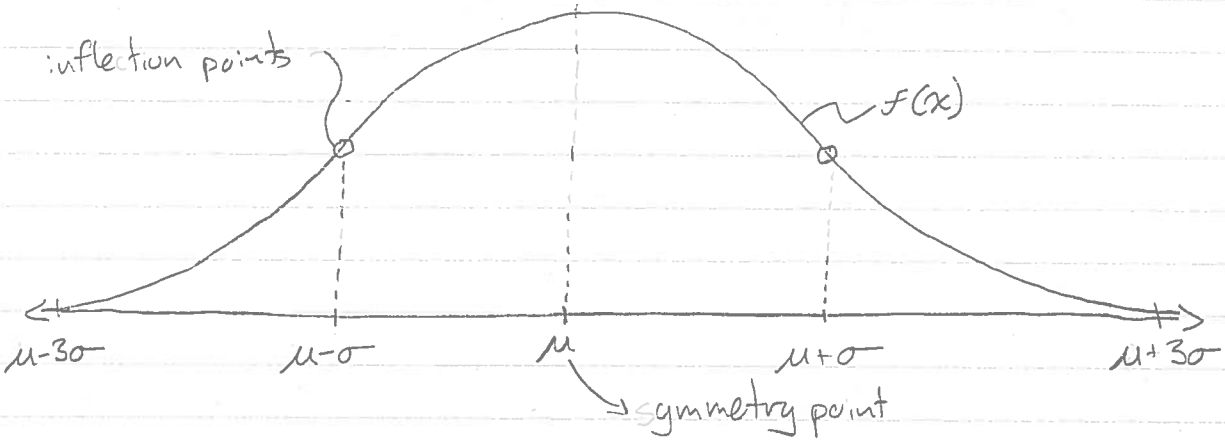
→ $P(a \le X \le b) = \int_a^b f(x)\,dx$ = area under the graph.

→ $P(X \le x) = P(X < x)$

7.2 Normal Distribution

- aka: normal random variables
- this is the most commonly encountered example of a continuous random variable.
- we say X has a normal distribution with parameters μ and $\sigma > 0$ if its density function is:

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



Area b/t $(\mu - 3\sigma)$ and $(\mu + 3\sigma)$ is 0.997

Interpretation of parameters:

$E(X) = \mu$ Recall $E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$

$Var(X) = \sigma^2$

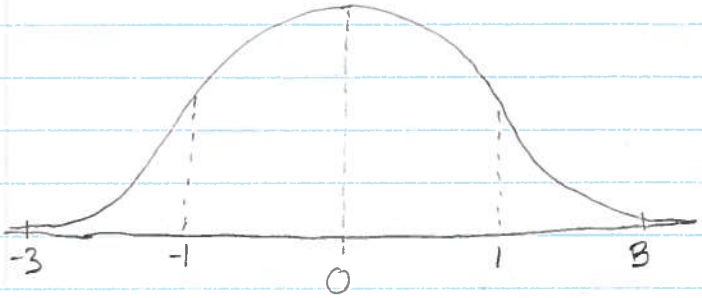
$\Rightarrow \mu$ is giving an indication about the average value of X and σ about the variability of X

Standard Normal Distribution

- This is the particular case $\{\mu = 0, \sigma = 1$
- The density of X is:

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

- Such a normal random variable is usually denoted by Z



Fundamental mathematical problem: there is no explicit formula for integrals of the form

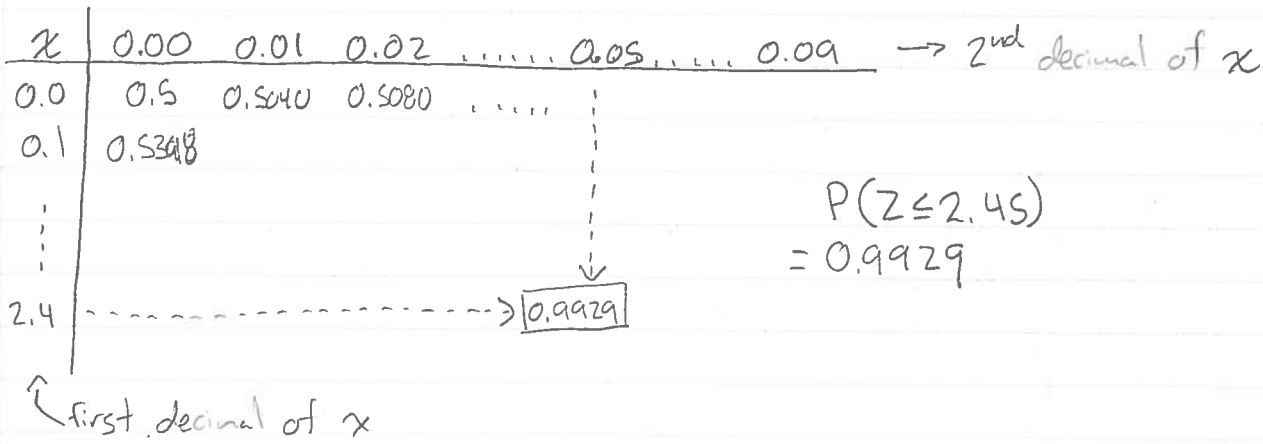
$$\underbrace{\int_a^b f(x)\,dx}_{P(a \le Z \le b)} = \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

Good news: we can still work with these probabilities, we know what they are. They can be computed using Table 17.2 and Table 17.3 or using R

$$P(a \le Z \le b) = P(Z \le b) - P(Z \le a)$$

Table 17.2 gives $P(Z \le x)$ for particular values $x < 0$
Table 17.3 gives $P(Z \le x)$ for particular values $x > 0$

Table 17.3



$P(Z \leq 2.45) = 0.9929$

Practice with the Tables (direct reading)

a) $P(Z \leq -1.52) = 0.0643$
$= P(Z \geq 1.52)$
$= 1 - P(Z \leq 1.52)$

b) $P(Z \leq 1.37) = 0.9147$
Table 17.3
↳ row 1.3
↳ column 0.07

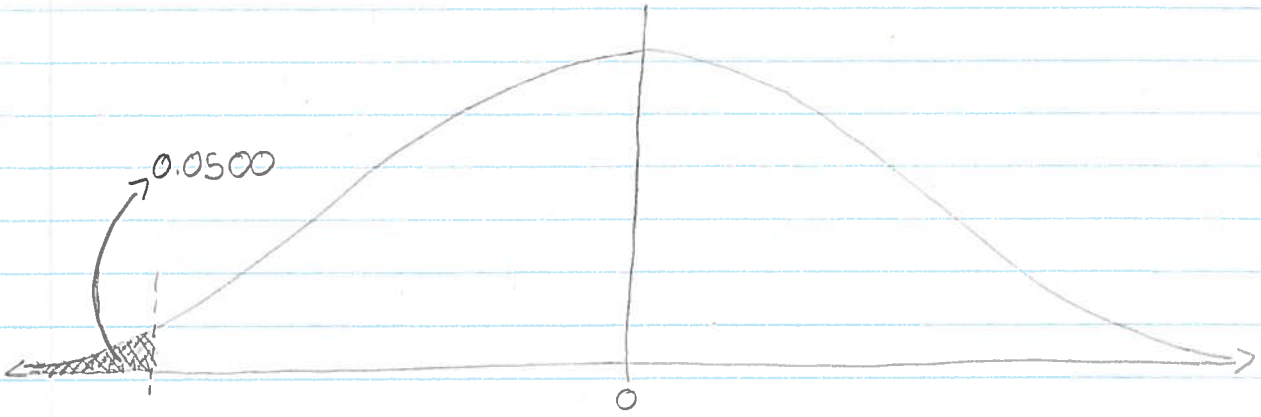
c) $P(Z \geq 1.98)$
$= 1 - P(Z \leq 1.98)$
$= 1 - 0.9761$
$= 0.0239$

d) $P(-1.21 \leq Z \leq 1.73)$
$= P(Z \leq 1.73) - P(Z \leq -1.21)$
$= 0.9582 - 0.1131$
$= 0.8451$

e) $P(Z = 1.50) = 0$

Indirect Reading in Tables 17.2, 17.3

a) Find the value z such that $P(Z \leq z) = 0.05$



look for 0.0500 inside table 17.2 :

$-1.64 \rightarrow 0.0505$

$-1.65 \rightarrow 0.0495$

(Z) (z)

so $z = -1.645$

MAT2379A 02/10/2013

7.2 - Normal Random Variables

We say that a random variable Z has a standard normal distribution if it is a continuous random variable with density function given by:

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

This means that $P(a < Z < b)$ for some values $a, b = \int_a^b f(x)\,dx$

Tables 17.2 and 17.3 give us the values

$$P(Z < a) = \int_{-\infty}^{a} f(x)\,dx$$

Ex (inverse reading): Find the point z such that $P(Z \leq z) = 0.75$

Since $0.75 > 0.5$, therefore $z > 0$

$$\left.\begin{array}{l} P(Z \leq 0.67) = 0.7486 \\ P(Z \leq 0.68) = 0.7517 \end{array}\right\} z \approx 0.675 \quad \text{(use the midpoint)}$$

Comanands in R to compute $P(Z \leq z)$ ← given

pnorm(z, 0, 1)

- 0 → mean μ of Z
- 1 → stand. deviation σ of Z

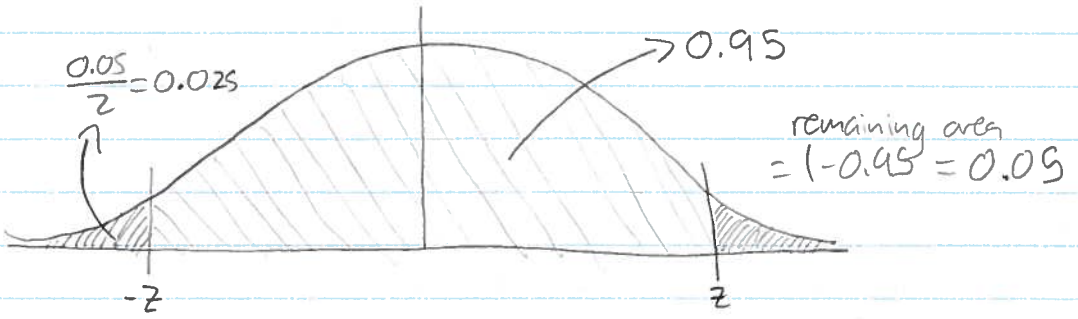
For X which is normal (μ, σ^2), we type:
pnorm(z, μ, σ)

for the inverse reading, suppose that we have to find the value z and we are given $P(Z \leq z) = p$ (p is given)
qnorm(p, μ, σ), $\mu = 0$, $\sigma = 1^2$

"q" → p-th quantile of Z

Ex) Find the value z such that

$$P(-z \leq Z \leq z) = 0.95$$



$P(Z \leq z) = 0.95 + 0.025 = 0.975$

$z = 1.96$ → look inside the table.

Non-Standard Normal Random Variables

aka = standard deviation procedure: if X is a Normal random variable with mean μ, variance σ^2, then

$Z = \frac{X - \mu}{\sigma}$ has a standard normal distribution.

Similar to Assign. 3

Example 4 The length of a fish is a normal random variable with mean $\mu = 54$ mm, variance $\sigma^2 = (4.5)^2 mm^2$. Hence the standard deviation is $\sigma = \sqrt{4.5^2} = 4.5$ mm

a) What is the probability that a randomly chosen fish is less than 60 mm long?

X = length of the fish; X is normal ($\mu = 54$, $\sigma = 4.5$)

$P(X < 60) = ?$

To compute this, there are 2 methods:

1) use R, type pnorm(60, 54, 4.5)
2) use Tables 17.2, 17.3

To use the tables, we need to standardize X.
We know that

$$\frac{X-\mu}{\sigma} = \frac{X-54}{4.5} = Z \text{ which is normal } (0,1)$$

$$P(X<60) = P\left(\frac{X-54}{4.5} < \frac{60-54}{4.5}\right)$$
$$= P(Z<1.33)$$
$$= 0.9082 \longrightarrow \text{direct reading from table 17.3}$$
$$\hookrightarrow \text{using R: pnorm(1.33, 0, 1)}$$

b) What is the probability that a fish is more than 51 mm long?

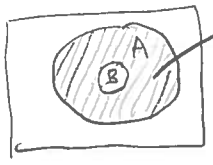
$$P(X \geq 51) = 1 - P(X \leq 51)$$
$$= 1 - P\left(X \leq \frac{51-54}{4.5}\right)$$
$$= 1 - P(X \leq -0.67)$$
$$= 1 - 0.2514$$
$$= 0.7486$$

in R: 1-pnorm(51, 54, 4.5)

c) What is the probability that a fish is less than 62mm long but more than 55mm long.

$$P(55 < X < 62)$$
$$= P(X<62) - P(X<55)$$
$$= P\left(Z < \frac{62-54}{4.5}\right) - P\left(Z < \frac{55-54}{4.5}\right)$$
$$= P(Z<1.78) - P(Z<0.22)$$
$$= (0.9625) - (0.5871)$$
$$= 0.3754$$

Using R:
pnorm(62, 54, 4.5) - pnorm(55, 54, 4.5)



$A \cap B' = A - B$

d) Find the length x_0 such that 20% of these fish have a length smaller than x_0.

$$P(X < x_0) = 0.20$$
$$P\left(Z < \frac{x_0-54}{4.5}\right) = 0.20$$

$$\frac{x_0-54}{4.5} = -0.845$$
$$x_0 = (-0.845)(4.5) + 54$$
$$= 50.1975 \text{ mm}$$

Using R: qnorm(0.20, 54, 4.5)

$$0.20 = P(X < x_0) = P\left(\underbrace{\frac{X-54}{4.5}}_{Z} < \underbrace{\frac{x_0-54}{4.5}}_{z_0}\right)$$

$$0.20 = P(Z < z_0)$$

$$z_0 = -0.845$$

★ look for 0.2000 ★ inside the table

$$\frac{x_0-54}{4.5} = -0.845$$

$$x_0 = 4.5(-0.845)+54$$

$$= 50.1975$$

since 0.2005 is very close to 0.2000 and 0.1977 is not, don't take the midpoint.

MAT 2379 A 07/10/2013

Midterm: Oct. 23 (in class)
Review: Oct. 21 (last year's exam)

Midterm is closed book
Bring: TI30 or Casio calculator

Material for midterm: up to chapter 7 (inclusively)

Chapter 9 Intro to Stats

9.1 Random Sampling and Data Description

What is a statistical problem? 3 elements characterize a statistical problem:

1. Population
ex: we are interested in factors related to heart disease in the population of people of age 18+

2. Variables (X, Y, Z, etc)
ex: age (numeric var.)
weight (numeric var.)
smoking status (yes or no)
family history (yes or no)

3. Sample
ex: we select randomly a group of n=50 persons in the population

Suppose for variable X (weight), we have 50 measurements corresponding to X, denoted by:
$x_1, x_2, x_3, \ldots, x_{50}$ → sample.

In this section, we will address the following problems:
a) what graphical methods can we use for analyzing a sample?
b) about what value does the variable fluctuate?
c) what is the variation in the data?

There are two types of variables:

1. Categorical variables

These take values which fall in several categories.
ex: smoking status (2 categories: yes, no)
blood type (4 categories: A, B, O, AB)
sex (2 categories: male, female)
colour of a flower (3 categories: red, white, pink)

2. Quantitative variables

These take numeric values. They can be:
a) discrete variables
b) continuous variables

Example 1)

15 newborns ($n = 15$ sample size)
Population: all newborns born in this hospital last year
{X = birth weight
{Y = mother's blood type

Part 1: Frequency Distributions

There are various graphical methods for summarizing a data set:

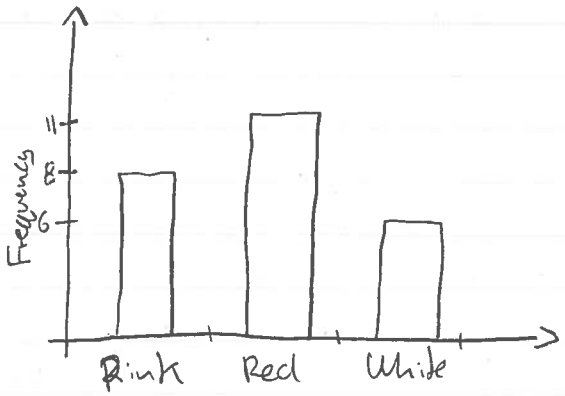
a) The Bar Chart

- used for categorical variables
- we draw a vertical bar whose height is equal to the frequency of the category
- frequency = number of observations in the sample which fall in a category.
- relative frequency = frequency / n

Example 1

(colour of poinsettias) 25 flowers:

colour	frequency	Related frequency
Pink	8	$8/25 = 0.32$
Red	11	$11/25 = 0.44$
White	6	$6/25 = 0.24$
Total	n=25	1.00



R-Code:

1) create a variable x with values 8, 11, 6:

```
> x = c(8, 11, 6)
```

2) create a variable y with values pink, red, white:

```
> y = c("pink", "red", "white")
```

3) create the bar chart:

```
> barplot(x, names.arg = y)
```

or

```
> barplot(x, names.arg = y, ylab = "Frequency",
          cex.axis = 1.5, cex.names = 1.5, cex.lab = 1.5)
```

b) Histograms
- a graphical method used for quantitative variables
- vertical bars has a height given by the frequency or relative frequency or probability density.
- the bars have the same width
- there is no space between the bars.

b.1) Histograms for discrete variables

Example 2 (piglets)

A company who owns a large number of pig farms is interested in the distribution of the number X of surviving piglets per sow.

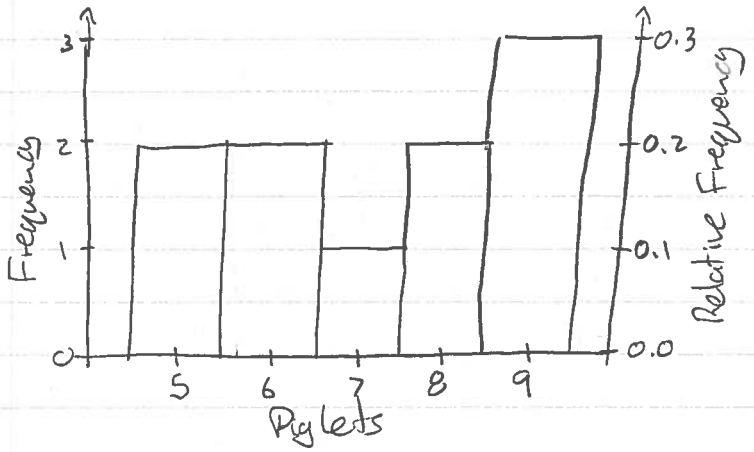
X is a discrete variable with values: $0, 1, 2, 3, \ldots, 20$

A sample of $n=10$ sows is selected. The observed values of X for this sample:

i	1	2	3	4	5	6	7	8	9	10
x_i	5	8	6	5	9	7	6	9	9	8

Here is the table of frequencies and relative frequencies.

Number of Survivors	Frequency	Relative Frequency
5	2	$2/10 = 0.2$
6	2	$2/10 = 0.2$
7	1	$1/10 = 0.1$
8	2	$2/10 = 0.2$
9	3	$3/10 = 0.3$
Total	$n=10$	1.00



R-Code

create the data set in R

```
> x = c(5, 8, 6, 5, 9, 7, 6, 9, 9, 8)
> hist(x, break = c(4.5, 5.5, 6.5, 7.5, 8.5, 9.5))
```

optional arguments:

```
xlab = "piglets"
ylab = "frequency"
main = "histogram"
```

b.2) Histograms for continuous variables

We arrange data into groups called bins and count how many data points fall into each bin.

Delicate issue: how many bins? $\sqrt{n}$ is usually selected but there is no general rule.

Example 3

Height of 15 students:

66.5 61.2 63.9 62.7 65.1 68.7 64.3 73.3 69.3
66.5 70.1 71.3 68.1 67.4 66.7

smallest: 61.2
largest: 73.3

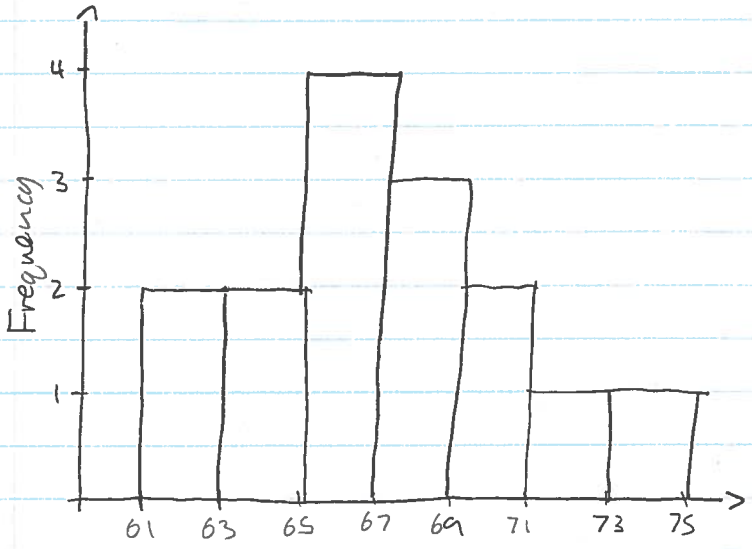
We will draw the histogram with 7 bins

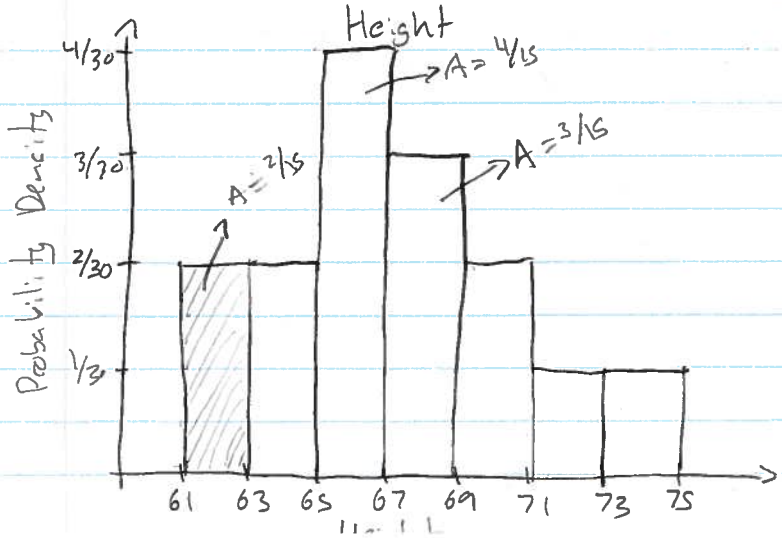
Bins:

Height	Frequency	Related Freq	Probability Density
61-63	2	2/15	(2/15)/2
63-65	2	2/15	(2/15)/2
65-67	4	4/15	(4/15)/2
67-69	3	3/15	(3/15)/2
69-71	2	2/15	(2/15)/2
71-73	1	1/15	(1/15)/2
73-75	1	1/15	(1/15)/2
Total	n=15	1.00	

Probability density = relative frequency ÷ bin width

Histogram of frequencies:





The area of a bar in a probability density graph is the relative frequency. Total area of all bars = 1

MAT2379 09/10/2013

To get data files, visit course website and click on "Data Files" on left-hand navigation menu. Right click on a data file and select "save target as". Save the .csv file anywhere.

* spaces in heading become full stops (periods)

To import a data file:
> table = read.table(file.choose(), header = TRUE, sep = "\t")
hit enter and browse to your data file; select it and click open.
To check if import was completed, enter the variable name:
> table → imported data will be displayed.
Rename the data set:
> x = table$density.of.earth
Generate a histogram:
> hist(x)
To add a curve:
> hist(x, prob = TRUE)

Section 9.1
Part 2: Descriptive Statistics

X = quantitative (ie numeric) variable
Observed values: $x_1, x_2, x_3 \ldots x_n$
The most commonly used descriptive statistics are:
a) the mean
b) the median
c) the minimum and maximum values
d) quartiles (IQR = interquartile range)
e) standard deviation

A) Measures of the center of data = mean and median

- The sample mean (or the average) is defined:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

R-Code = mean(x)

Example 1 X = wieght (in kg) of a 12-year old girl

n = 10 girls

data: $x_1 = 27.7, x_2 = 31.5, x_3 = 30.9, x_4 = 29.6, x_5 = 27.0$

$x_6 = 38.1, x_7 = 32.4, x_8 = 31.1, x_9 = 36.7, x_{10} = 28.4$

$$\bar{x} = \frac{1}{10}(27.7 + \ldots + 28.4) = 31.34$$

∴ 31.134 is interpreted as an approximation for $\mu = E(x)$

i^{th} deviation $= x_i - \bar{x}$

for $i=1$, $x_1 - \bar{x} = 27.7 - 31.34 < 0$

The sum of all deviation in a sample is:

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

Why? Because $\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = n\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x}\right)}_{0} = 0$

The median $\tilde{x}$ is a value which devides a sample into two groups of the same size; half of the datapoints are smaller than $\tilde{x}$ and half of the data points are greater than $\tilde{x}$. To define $\tilde{x}$ we need to arrange the data in ingreasing order: $y_1 \leq y_2 \leq 3 \leq \ldots \leq y_n$

$$\tilde{x} = \begin{cases} y_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(y_{\frac{n}{2}} + y_{\frac{n}{2}+1}\right) & \text{if } n \text{ is even} \end{cases}$$

Ex) if $n=7$ then $\tilde{x} = y_{\frac{3+4}{2}}$

y_1 y_2 y_3 $\boxed{y_4}$ y_5 y_6 y_7

3 values ← middle point → 3 values

Ex) if $n=6$ then $\tilde{x} =$

y_1 y_2 $\underbrace{y_3 \quad y_4}$ y_5 y_6

average of these two middle values.

Example 1 Cont. Find the median of this data.

$y_1 = 27.0$ $y_2 = 27.7$ $y_3 = 28.4$ $y_4 = 29.6$ $y_5 = 30.0$
$y_6 = 31.1$ $y_7 = 31.5$ $y_8 = 32.4$ $y_9 = 36.7$ $y_{10} = 38.1$

$n=10 \rightarrow$ even $\quad n/2 = 5 \quad n/2 + 1 = 6$

$$\tilde{x} = \frac{1}{2}(y_5 + y_6) = \frac{1}{2}(30.9 + 31.1) = 31.0$$

{ 5 values in the sample are < 31.0
{ 5 values in the sample are > 31.0

R-Code for median = median (x)

Example 2

x = systolic blood pressure after 10 minutes of cardio exercise
The data for a sample of size $n=5$:
$x_1 = 105$, $x_2 = 120$, $x_3 = 111$, $x_4 = 125$, $x_5 = 132$
Find the mean and median.

$$\bar{x} = \frac{105 + 120 + \ldots + 132}{5} = 118.6$$

For median, arrange data as:
$y_1 = 105$. $y_2 = 111$ $y_3 = 120$ $y_4 = 125$ $y_5 = 132$

$n = 5 \rightarrow$ odd $\quad \tilde{x} = y_{\frac{n+1}{2}} = y_{6/2} = y_3 = 120$

Remark: median is "robust", ie: it is not affected by a small change in the data.
(in ex3, replace $y_1 = 105$ by 93, the median remains the same)
The mean is "efficient", ie it uses all the information.

B) Boxplots

Quartiles are values that divide the data into 4 equally sized groups. They are denoted by: $q_1, q_2 = \tilde{x}, q_3$.
The first quartile is defined by:

$$q_1 = \begin{cases} \text{i) } y_{\left(\frac{n+1}{4}\right)} & \text{if } \frac{n+1}{4} \text{ is an integer } (\in \mathbb{Z}) \\ \left.\begin{array}{ll} \text{ii) } 0.75y_r + 0.25y_{r+1} & \text{if } \frac{n+1}{4} = r + \frac{1}{4} \\ \text{iii) } 0.5y_r + 0.5y_{r+1} & \text{if } \frac{n+1}{4} = r + \frac{2}{4} \\ \text{iv) } 0.25y_r + 0.75y_{r+1} & \text{if } \frac{n+1}{4} = r + \frac{3}{4} \end{array}\right\} r \in \mathbb{Z} \end{cases}$$

ex for i) $n = 7$, then $\frac{n+1}{4} = 8/4 = 2$

y_1 $\underbrace{y_2}_{\substack{\text{first} \\ \text{quartile} \\ q_1}}$ y_3 $\underbrace{y_4}_{\tilde{x}}$ y_5 $\underbrace{y_6}_{\substack{\text{third} \\ \text{quartile} \\ q_3}}$ y_7

ex for ii) $n = 8$, then $\frac{n+1}{4} = 9/4 = 2 + 1/4 = 2.25$
2.25 lies between 2 and 3, so q_1 also lies b/t y_2 and y_3
$2.25 = (0.75) \times 2 + (0.25) \times 3$
$q_1 = (0.75)y_2 + (0.25)y_3$

Example 1 Cont.

Find the first quartile q_1 for the data set.

Arrange the data in increasing order

$n=10 \quad \frac{n+1}{4} = \frac{11}{4} = 2.75 = 2 + \frac{3}{4}$

2.75 is b/t 2 and 3 and is closer to 3

So, q_1 should be b/t y_2 an y_3 but closer to y_3.

We assign the large weight (0.75) to y_3 and the smaller weight (0.25) to y_2.

$$q_1 = (0.25)y_2 + (0.75)y_3$$
$$= (0.25)(27.7) + (0.75)(28.4)$$
$$= 28.225$$

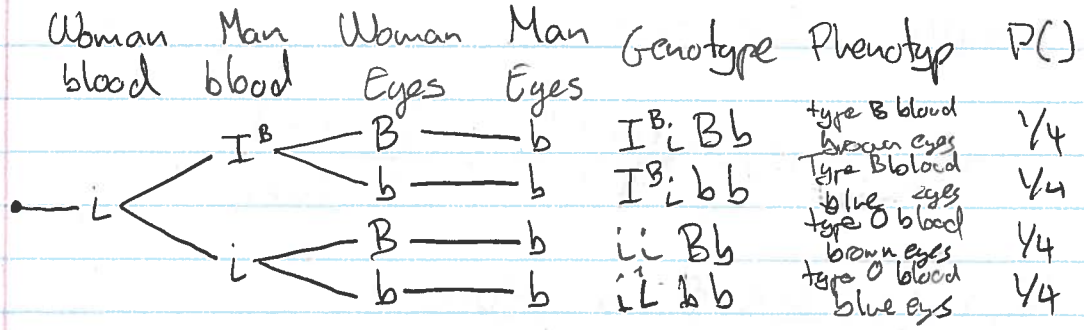
MAT2379 A 21/10/2013

Review

1. Type A blood: $I^A I^A$ or $I^A i$ Brown (B)
   Type B blood: $I^B I^B$ or $I^B i$ Blue (b)
   Type AB blood: $I^A I^B$
   Type O blood: ii

woman → blood type O (ii)
      → brown eyes (Bb)
man → blood type B ($I^B i$)
    → blue eyes (bb)

P(child has brown eyes and blood type O) = ?



P(ii Bb) = 1/4 (A)

2. P(A) = P(B) = 0.2
   A and B are mutually exclusive events, $(A \cap B) = \emptyset$
   ↳ A and B cannot occur at the same time

$P(A' \cap B')$
= complement of $A \cup B$
= $1 - P(A \cup B)$
= $1 - (P(A) + P(B))$
= $1 - (0.2) - (0.2)$
= 0.6 (D)

3. 2009 patients
   ↳ 1000 have a smoking history
   ↳ 1500 have respiratory complications
   ↳ 90% of patients with smoking history, have resp. complications

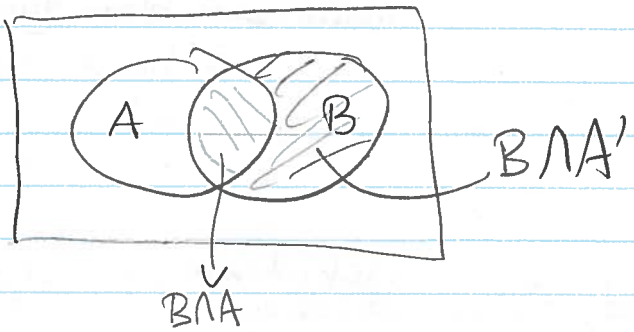
A = smoking history; $P(A) = 1000/2009$

B = respiratory complications; $P(B) = 1500/2009$

$B|A$ = patients with a smoking history have resp. compl.

$P(B|A) = 0.90$

$$\text{relative risk} = \frac{P(B|A)}{P(B|A')}$$



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

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

$$0.90 = P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(B \cap A)}{1000/2009}$$

$$P(B \cap A) = \frac{1000}{2009} \times 0.9 = \frac{900}{2009}$$

Hence: $P(B \cap A') = \frac{1500}{2009} - \frac{900}{2009} = \frac{600}{2009}$

$$P(B|A') = \frac{600/2009}{1009/2009} = \frac{600}{1009}$$

$$\text{relative risk} = \frac{0.9}{600/1009} = 1.5135$$

$$= \frac{P(\text{complications} \mid \text{smoking})}{P(\text{complications} \mid \text{non smoking})} > 1$$

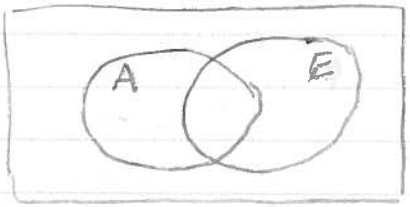
(A)

4. Drug A is efficient; Prob = 0.75
Drug B is efficient; Prob = 0.55

50% are treated with drug A
50% are treated with drug B

If the drug is efficient for a certain patient, what is the probability that the patient was treated with drug A?

A = treated with drug A
A' = treated with drug B
$P(A) = P(A') = 0.5$
E = drug is effecient
$P(E|A) = 0.75$
$P(E|A') = 0.55$



$P(A|E) = ?$

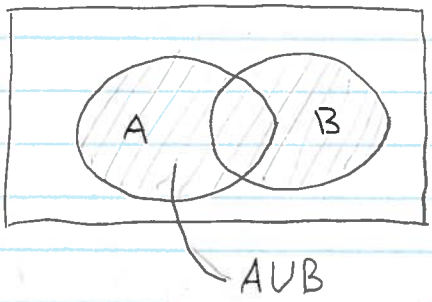
$$P(A|E) = \frac{P(A \cap E)}{P(E)}$$

$$P(A \cap E) = P(A)P(E|A) = (0.5)(0.75) = 0.375$$

$$P(E) = P(E \cap A) + P(E \cap A')$$
$$= P(A)P(E|A) + P(A')P(E|A')$$
$$= (0.5)(0.75) + (0.5)(0.55)$$
$$= 0.65$$

$$P(A|E) = \frac{P(A \cap E)}{P(E)} = \frac{(0.375)}{(0.65)} = 0.5769$$ (E)

5. T_1, T_2 are independent tests
T_1 is incorrect 5% of patients
T_2 is incorrect 2% of patients



$A = T_1$ is correct; $P(A) = 0.95$
$B = T_2$ is correct; $P(B) = 0.98$

Note: independence means
$P(A \cap B) = P(A)P(B)$
$= (0.95)(0.98)$
$= 0.931$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
$= (0.95) + (0.98) - (0.931)$
$= 0.999$

B

6.

x	0	1	2
$P(X=x)$	0.5	0.3	0.2

$E(X) = \sum_x xf(x) = 0.7$

$Var(X) = \sum(x-\mu^2)f(x) = 0.61$

A

7. Offspring has a prob. of 0.25 of being an albino. Couple has 5 children. What is the probability that exactly 2 children are albino?

X = number of children that are albino; random variable
$x = 0, 1, 2, 3, 4, 5$

$n = 5$
$p = 0.25$
} binomial random variable X

$P(X=2) = \binom{5}{2}(0.25)^2(1-0.25)^{5-2} = 0.264$ (D)

8. Average waiting time for surgery is 17.3 weeks. Assuming that the waiting time is normally distributed with std. dev. 1.9 weeks. What is the prob. that a randomly chosen person has to wait more than 22 weeks?

X = waiting time with normal distribution with
$\mu = 17.3$
$\sigma^2 = 1.9$

$P(X > 22)$
$= 1 - P(X \leq 22)$
$= 1 - P\left(Z \leq \frac{22-(17.3)}{(1.9)}\right)$
$= 1 - P(Z \leq 2.47)$
$= 1 - (0.9932)$
$= 0.0068$ (B)

9. X = diameter at breast height with normal distribution with
$\mu = 7.5$ m
$\sigma = 0.5$

$P(X > 7.72)$
$= 1 - P(X \leq 7.72)$
$= 1 - P\left(Z \leq \frac{7.72-(7.5)}{(0.5)}\right)$
$= 1 - P(Z \leq 0.44)$
$= 1 - (0.6700)$
$= 0.33$

$n = 5$, $p = 0.33$
Y = # of trees in a sample of 5 w diameter > 7.72

$P(Y=2) = \binom{5}{2}(0.33)^2(1-0.33)^{5-2} = 0.328$

(E)

10. Omitted for this exam!

11. x in R (heights of 100 students)

$\text{Var}(x) =$

Omitte for this exam!

12. X is binomial with $n = 100$, $p = 0.25$

$P(16 \leq X \leq 31) = ?$

$= P(X = 16, 17, 18, \ldots, 31)$

$= P(X \leq 31) - P(X \leq 15)$

$= (0.9306511) - (0.01108327)$

$= 0.9196$

(A)

MAT2379 28/10/2013

Final Exam
- december 10th at 9:30 (location TBD)

Section 9.1.2 = Descriptive Statistics

Quartiles: q_1 , $\tilde{x}$, q_3
1st quartile, median, 3rd quartile

$$\tilde{x} = \begin{cases} y_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ (1/2)(y_{n/2} + y_{n/2+1}) & \text{if } n \text{ is even} \end{cases}$$

$$q_3 = \begin{cases} y_r & \text{if } \frac{3(n+1)}{4} = r \\ 0.75y_r + 0.25y_{r+1} & \text{if } \frac{3(n+1)}{4} = r + 1/4 \\ 0.5y_r + 0.5y_{r+1} & \text{if } \frac{3(n+1)}{4} = r + 2/4 \\ 0.25y_r + 0.75y_{r+1} & \text{if } \frac{3(n+1)}{4} = r + 3/4 \end{cases}$$

ex: $n = 14$ $\quad \frac{3(n+1)}{4} = \frac{3(15)}{4} = \frac{45}{4} = 11 + \frac{1}{4} = 11.25$

∴ 11.25 falls b/t 11 and 12 but is closer to 11
∴ q_3 falls b/t y_{11} and y_{12} but is closer to y_{11}
$q_3 = 0.75y_{11} + 0.25y_{12}$

Example 1 cont.
$n = 10$

$y_1 = 27.0$	$y_2 = 27.7$	$y_3 = 29.4$	$y_4 = 29.6$	$y_5 = 30.9$
$y_6 = 31.1$	$y_7 = 31.5$	$y_8 = 32.4$	$y_9 = 36.7$	$y_{10} = 38.1$

Find the 3rd quartile
$n = 10$ $\quad \frac{3(10+1)}{4} = \frac{3(11)}{4} = \frac{33}{4} = 8 + 1/4 = 8.25$
∴ q_3 falls bt y_8 and y_9 but closer to y_8

$q_3 = 0.75y_8 + 0.25y_9$
$= 0.75(32.4) + 0.25(36.7)$
$= 33.475$

Recall that $q_1 = 28.225$; $\tilde{x} = 31.0$

The distance b/t q_1 and q_3 is called the inter-quartile range (IQR): $IQR = q_3 - q_1$

ex 1 cont.) IQR: $(33.475) - (28.225) = 5.25$

The range of the data is $R = y_n - y_1$

ex 1 cont.) $R = (38.1) - (27.0) = 11.1$

The 5-number Summary is: $(y_1, q_1, \tilde{x}, q_3, y_n)$

IQR

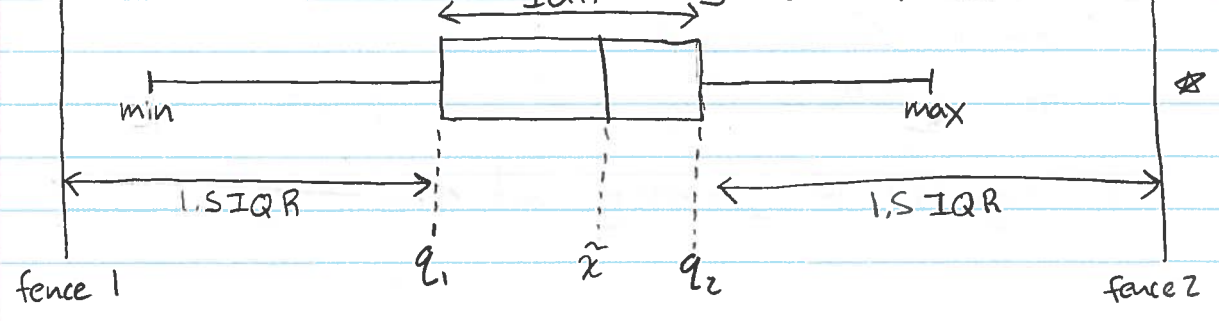
R

R-Code: >summary(x)

* note: the quartiles will be slightly different; to get our quartiles we type >quantile(x, type=6)

Boxplot

- graphical method for illustrating the 5-number summary.
- draw a box from q_1 to q_3
- draw a line where the median falls
- draw fences at a distance of 1.5 IQR from the two quartiles
- draw whiskers from the ends of the box, the smallest and largest values in the sample, within the 2 fences
- values outside the fences are denoted by * and are called outliers



Example 1 cont.: construct the boxplot by hand and see if there are any outliers.

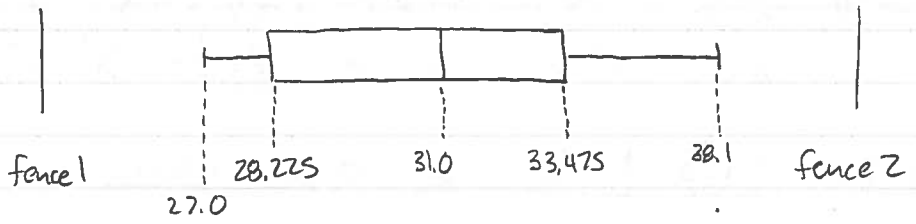
$IQR = 5.25$; $q_1 = 28.225$; $q_3 = 33.475$

Fence 1 $= q_1 - 1.5IQR = 28.225 - 1.5(5.25) = 20.35$
Fence 2 $= q_3 + 1.5IQR = 33.475 + 1.5(5.25) = 41.35$

Any values < 20.35 is an outlier
Any values > 41.35 is an outlier

Outliers: none!

Boxplot:



R-Code: >boxplot(x)

Measures of Dispersion

The sample variance is defined as: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Can also be computed by:
$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$

s^2 is a measure of variability;
s^2 large $\Rightarrow$ lots of variability in the data
s^2 small $\Rightarrow$ data is clustered around $\bar{x}$

Example 2
- systolic blood pressure
- $n = 5$; $x_1 = 105$, $x_2 = 120$, $x_3 = 111$, $x_4 = 125$, $x_5 = 132$

Find the sample variance:

$$\bar{x} = (1/5)(105 + 120 + 111 + 125 + 132) = 118.6$$

$$s^2 = \frac{1}{5-1}\left[105^2 + 120^5 + 11^2 + 125^2 + 132^2 - 5(118.6)^2\right] = 116.3 \text{ mmHg}^2$$

The square root of s^2 is called the sample standard deviation.

$$s = \sqrt{s^2} \quad ; \quad s = \sqrt{116.3} = 10.78 \text{ mmHg}$$

R-Code: > Var(x) gives the sample variance
> sd(x) gives the sample standard deviation

Recall that if X is a random variable then
→ the mean of X is $E(X) = \sum x P(X = x)$ if X is discrete
→ the variance of X is $Var(X) \overset{\text{all } x}{=} \sum_{\text{all } x} (x - E(X)^2) P(X = x)$

Section 9.1.3: transformations of variables

-a linear transformation of a sample which is called

$$x_1, x_2, \ldots, x_n$$

is a new sample obtained as follows:

$x_1' = ax_1 + b$ (a, b are constants)
$x_2' = ax_2 + b$ note: $\bar{x}' = a\bar{x} + b$
$\vdots$ $(s')^2 = a^2 s^2$
$x_n' = ax_n + b$ $s' = as$

Example (temperature)
In degrees Celsius we have the following data: $x_1 = 36.23$, $x_2 = 36.41$, $x_3 = 36.44$, $x_4 = 36.15$

$\bar{x} = 36.39$, $s^2 = 0.076$, $s = 0.2757$

We want to express the data in degrees Fahrenheit. IE, we use the linear transformation:

$X' = 1.8X + 32$; X' = fahrenheit, X = Celsius

The new data is:

$x_1' = 1.8(36.23) + 32 = 97.214$

$x_2' = 1.8(\ldots\ldots = 97.538$

$x_3' = 1.8 \quad = 98.186$

$x_4' = 1.8 \quad = 97.07$

$\overline{x'} = 1.8(36.39) + 32 = 97.502$

$s^2 = 1.8^2(0.076) = 0.246$

$s = 1.8(0.2752) =$

Logarithmic Transformations

- the logarithmic transformation of $x_1, x_2, \ldots x_n$ is a new sample given by:

$$y_1 = \ln(x_1),\ y_2 = \ln(x_2),\ \ldots,\ y_n = \ln(x_n)$$

- R-Code: if the data $x_1, x_2, \ldots, x_n$ is saved in x, then type:

```
>y=log(x)
```

- the sample mean of $y_1, y_2, \ldots, y_n$ (is geometric mean of $x_1, x_2, \ldots x_n$)

$$\bar{y} = (1/n)\sum_{i=n}^{n} y_i \qquad e^{\bar{y}}$$

- the sample std. deviation of $y_1, y_2, \ldots y_n$ is:

$$s_y = \sqrt{(1/n-1)\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

e^{s_y} is called the geometric std. dev. of $x_1, x_2, \ldots x_n$

Example 2 cont.

$x_1 = 105, x_2 = 120, x_3 = 111, x_4 = 125, x_5 = 132$

Find the geometric mean and the geometric std. dev.

$y_1 = \ln(105) = 4.65$
$y_2 = \ln(120) = 4.79$
$y_3 = \ln(111) = 4.71$
$y_4 = \ln(125) = 4.83$
$y_5 = \ln(132) = 4.88$

$\bar{y} = (1/5)(4.67 + 4.79 + 4.71 + 4.83 + 4.88)$
$= 4.77$

geometric mean: $e^{\bar{y}} = e^{4.77} = 118.21$

$s_y = 0.0915$

geometric std. dev.: $e^{s_y} = e^{0.0915} = 1.096$

R-Code:

```
> x = c(105, 120, 111, 125, 132)
> y = log(x)
> m = mean(y)   ~> calculates ȳ
> g = exp(m)    ~> calculates e^ȳ
> s = std(y)    ~> calcalates sy
> k = exp(s)    ~> calculates e^sy
```

MAT2379 30/10/2013

R-Code

importing data `>table = read.table (file.choose(), header = TRUE, sep = "\t")`

display data `>table`

assigning data
```
>x = table $ king
>y = table $ gentoo
>z = table $ chinstrap
```

5 number summary:
```
>summary (x)
```

Quartiles:
```
>quantile (x, type = 6)
```

Mean:
```
>mean (x)
```

St. Dev.
```
>sd (x)
```

Variance
```
>var (x)
```

Median
```
>median (x)
```

Histogram
```
>hist (x)
```

Transformation (natural log)
$y = \log(x)$

Boxplot
> boxplot(y)

Side-by-Side Boxplots
> boxplot(x~y)

Section 9.2 - Sampling Distributions and Point Estimations

So far,
- we learned how to calculate probabilities associated with random events: ch. 2, 3, 4, 5
- we looked at how to describe a measurement X (mean $\mu = E(X)$, variance $\sigma^2 = Var(X)$): ch. 6, 7
- we looked at how to analyze a data set $x_1, x_2, \ldots, x_n$

We are interested in drawing conclusions about the parameters μ and σ^2 which speak about X which is unknown, using a sample $x_1, x_2, \ldots x_n$ which are drawn from a population.

Example: X = growth of a randomly chosen seedling in a lab.
μ = mean growth of all seedlings in the lab
select a sample of $n = 6$ seedlings: $x_1 = 20$cm, $x_2 = 22$cm, $x_3 = 19$cm, $x_4 = 21$cm, $x_5 = 24$cm, $x_6 = 19$cm

We introduce some theoretical variables called $X_1, X_2, \ldots, X_6$

Interpretation: $x_1 = 20$ is one of the possible values of X_1, every time we repeat this experiment, we get new data $x_1, \ldots x_n$ but we use the same notation $X_1, X_2, \ldots X_n$ for the theoretical values.

$X_1, X_2, \ldots, X_n \rightarrow$ theoretical sample

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow$ theoretical sample mean

Ex cont: what is the observed value of $\overline{X}$ for this particalar data?

$$\overline{X} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}(20 + 22 + \ldots + 19) = 20.67$$

We assume that the theoretical values $X_1, X_2, \ldots, X_n$ are:
- independent of eachother
- they have the same distribution

Therefore, they all have the same mean: $E(X_1) = E(X_2) = \ldots = E(X_n) = \mu$
↳ population mean
They also have the same variance: $Var(X_1) = Var(X_2) = \ldots = Var(X_n) = \sigma^2$
↳ population variance

Theoretical Values	X_1	X_2	...	X_n	$\overline{X}$	S^2
Observed Values	x_1	x_2	...	x_n	$\overline{x}$	s^2

Population Parameters	μ	σ^2
Sample Estimators	$\overline{x}$	s^2

$\overline{X}$ is an estimator of μ, the observed value $\overline{x}$ is an approximation of μ.

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}(n\mu) = \mu$$

$$Var(\overline{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \sigma^2/n$$

- variability decreases with sample size; the larger the sample size n is, the smaller variability in $\overline{X}$

The standard deviation of $\bar{X}$ is:

$$\sqrt{Var(\bar{X})} = \sqrt{\sigma^2/n} = \sigma/\sqrt{n}$$

Example 2: We denote by X the weight (in kg) of a female lamb at birth. We assume that X has a normal distribution with mean $\mu = 6.2$ kg and $\sigma = 0.6$ kg. Consider a sample of $n = 49$ lambs. Let $\bar{X}$ be the mean of this sample. Compute the expectation of $\bar{X}$ and the standard deviation of $\bar{X}$

$E(\bar{X}) = \mu = \underline{6.2 \text{ kg}}$
$Var(\bar{X}) = \sigma^2/n = (0.6)^2/(49)$
$\sqrt{Var(\bar{X})} = \sqrt{\sigma^2/n} = \sigma/\sqrt{n} = (0.6)/\sqrt{(49)} = 0.6/7 = \underline{0.0857}$

Note that there is a decrease in the std. dev. from 0.6 to 0.0857

The Distribution of $\bar{X}$

- central limit theorem
- if the sample size n is large enough, then

$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$ has approximatly a normal distribution. with mean $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$

Illustration with R of the C.L.T. will be as follows:

1) ask computer to give a sample (generate) of size n from binomial $(m = 5, p = 0.3)$

> rbinom(n, m, p)

- compute the mean $\bar{x}_1$ of this sample.

2) generate another sample of the same size from the same binomial $(m = 5, p = 0.3)$ and compute the mean $\bar{x}_2$ of this new sample.

3) repeat this procedure 5000 times

At end, you have $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{5000}$

Another example: Binomial($n=1, p=0.5$), the first sample of size $n=100$ will look like $\{1,0,0,1,\ldots,1,0,\ldots,0,1,1\}$ (100 times).

4) draw the histogram of all 5 values
The C.L.T says that this histogram should resemble a normal curve/density.

R-code: summary of R instructions (Part 2)

Example 3: a botanist has planted tomatoe seedlings and is measuring their growth after 30 days. Let X be the growth of a randomly chosen seedling. Assume that X is a continuous random variable with mean $\mu = 10$ and standard deviation $\sigma = 3.5$. The botanist is selecting a sample of $n=64$ seedlings. Let $\bar{X}$ be the mean of this sample. Give an approximation for the probability that $\bar{X}$ is greater than 11.2.

$\bar{X}$ has approximate normal distribution with

$\mu_{\bar{X}} = \mu = 10 \quad \sigma_{\bar{X}} = \sigma/\sqrt{n} = 3.5/\sqrt{64} = 0.4375$

$$P(\bar{X} > 11.2) = 1 - P\left(Z < \frac{11.2-10}{0.4375}\right) = 1 - P(Z < 2.74) = \underline{0.0031}$$

MAT2379 04/11/2013

9.3 - Assessing Normality

Question: can we assume that the data $x_1, x_2, \ldots, x_n$ is normally distributed?

In step 1, we look at the histogram of the data. If the histogram does not have a symmetric bell-shape, then the data is not normally distributed.

If the histogram seems to be symmetric and bell-shaped then we need more advanced methods, namely QQ-plots. Q = quantile.

How do we construct a QQ plot?

- assume that $X_1, X_2, \ldots, X_n$ are independent random measurements that are normally distributed with mean μ and variance σ^2?

By standardization, for each $i = 1, \ldots, n$

$$Z_i = \frac{X_i - \mu}{\sigma} \quad \text{has a } N(0,1) \text{ distribution}$$

So, $X_i = \mu + \sigma Z_i$ for each $i = 1, \ldots, n$
Take expectation on both sides

$$\underbrace{E(X_i)}_{\mu_X = \mu} = \mu + \sigma \underbrace{E(Z_i)}_{\mu_Z}$$

$$\mu_X = \mu + \sigma \mu_Z \qquad \mu_Z = 0$$

we arrive at the conclusion: $\mu_X = \mu$, this is nothing new!

We need to do something more complicated. We will arrange the values $X_1, X_2, \ldots, X_n$ in increasing order as: $Y_1 < Y_2 < \ldots < Y_n$

We will also arrange the $Z_1, Z_2, \ldots, Z_n$ as: $W_1 < W_2 < \ldots < W_n$

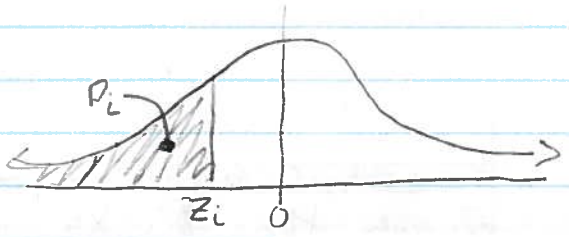
Hence we have: $Y_i = \mu + \sigma W_i$; now take the expectation

$$E(Y_i) = \mu + \sigma \underbrace{E(W_i)}_{z_i}$$

z_i is called a normal score
z_i can be computed using tables 17.2, 17.3, since

$$P(Z \leq z_i) = p_i$$

where: $p_i = \frac{i - 3/8}{n + 1/4}$



QQ-plot
- the plot of the pairs $(y_1, z_1), (y_2, z_2), \ldots, (y_n, z_n)$
- if the plot of these n points seem to be linear, then we say that the data $x_1, x_2, \ldots x_n$ seem to be normally distributed.
- the line which best fits the points $(y_1, z_1), \ldots, (y_n, z_n)$ is the line:

$$y = \hat{\mu} + \hat{\sigma} z \quad \text{where } \hat{\mu} = \bar{x}, \ \hat{\sigma} = s$$

Example: X = weight loss (in lbs) of one person who participated in a weightloss program. This person is chosen from seven participants. $X_1, X_2, \ldots, X_n$ are the random weight losses of 7 persons in this program. We have the following observed data: $x_1 = 7.06$, $x_2 = -0.61$, $x_3 = 3.87$, $x_4 = 3.73$, $x_5 = 3.61$, $x_6 = -2.14$, $x_7 = -5.28$. Is it reasonable to assume that this data comes from a normal distribution. (ie. is it reasonable to assume that X has a normal distribution)

Solution: we'll draw a QQ plot.

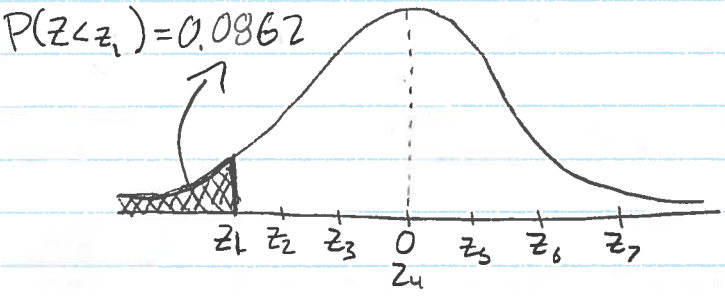
$y_1 = -5.28$ $y_2 = -2.14$ $y_3 = 0.61$ $y_4 = 3.61$

$y_5 = 3.73$ $y_6 = 3.87$ $y_7 = 7.06$

we also have to find the normal scores: $z_1, z_2, \ldots z_n$

for z_1:

$$P_1 = \frac{i - 3/8}{n + 1/4} = \frac{1 - 3/8}{7 + 1/4} = \frac{5/8}{29/4} = 0.0862$$



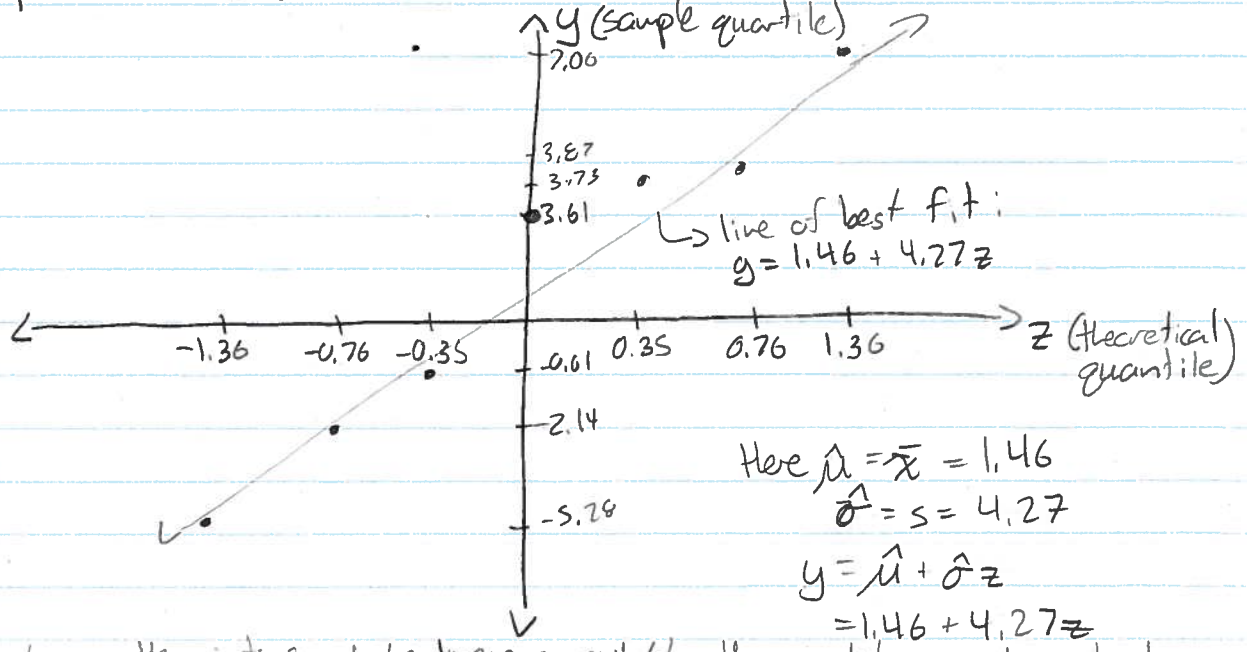
Look inside table 17.2 (inverse reading) to find $z_1 = -1.36$. Do the save for all values

$z_1 = -1.36$ $z_2 = -0.76$ $z_3 =$ $z_4 = 0$

$z_5 =$ $z_6 =$ $z_7 =$

We need to plot the pairs (y_i, z_i), $i = 1, \ldots, 7$

y_i	-5.28	-2.14	0.61	3.61	3.73	3.87	7.06
z_i	-1.36	-0.76	-0.35	0	0.35	0.76	1.36



Here $\hat{\mu} = \bar{x} = 1.46$

$\hat{\sigma} = s = 4.27$

$y = \hat{\mu} + \hat{\sigma} z$

$= 1.46 + 4.27z$

Conclusion = the points seem to be linear, so probably, the normality assumption about x is satisfied

Comment = in the textbook, the QQ plots were done with a different program (MINITAB) and theoretical quantiles (z_i) are on the vertical axis and sample quantiles (y_i) are on the horizontal axis.

R Software

Central Limit Theory

>rbinom(12, 1, 0.5) generates a list of 0's and 1's
>rnorm(3, 0, 1) generate a sample from the normal distribution.

⊗ program not needed to be memorized.

```
>p=ppoints(n)
>z=qnorm(p, 0, 1)
>x=
```

```
>qqnorm(x)
>abline(mean(x), sd(x))
```

MAT2379 06/11/2013

Assignment 4
6d. use R-Code sqrt(x) to get the square root

Chapter 10 - Confidence Intervals

Goal = estimate a parameter

Example parameters:
1) mean μ
2) proportion p of individuals in a pop. who have a certain characteristic

Conclusions will be based on a sample

X = measurement of interest, $\mu = E(X)$, $\sigma^2 = Var(X)$
$x_1, \dots, x_n$ are the observed values of X
$X_1, \dots, X_n$ are the theoretical values

- A point estimator of the population mean μ is:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the observed value of X and is called an estimate of μ (μ unknown)

- A point estimater of the population variance σ^2 is:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

The sample variance:

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

is the observed value of S^2 and is called an estimate of σ^2

Summary

Parameters	(theoretical) Estimator	(numeric) Estimate
μ	$\bar{X}$	$\bar{x}$
σ^2	S^2	s^2

Example 1 X = weight of a randomly chosen polar bear cub at birth. Set μ be the average of X (μ is unknown). Set σ^2 be the variance of X. We have the following data for a sample of 5:

$x_1 = 785 \quad x_2 = 825 \quad x_3 = 671 \quad x_4 = 981 \quad x_5 = 732$

If we compute the sample mean:

$\bar{x} = (1/5)(785 + 825 + 671 + 981 + 732) = 798.8$

$\therefore$ 798.8 g is an estimate (approximation) for the average weight of a randomly chosen polar bear cub at birth μ.

The sample standard deviation is:

$s = \sqrt{(1/4)(785^2 + \ldots + 732^2 - 5(798.8)^2)}$

$= \sqrt{13\,717.2}$

$= 117.1205$ g

s is an estimate (ie an approximation) of σ, which is the standard deviation of the weight of a polar bear cub (for the entire population)

10.1 - Confidence Intervals for μ

- case when σ^2 is known

Idea: we are not happy with one (single) value $\bar{x}$ as an estimate for μ. Instead of this, we would like to provide a range of values $[L_1, L_2]$ such that

$$P(L_1 \leq \mu \leq L_2) = 0.95 \text{ or } 0.99 \text{ or } 0.90$$

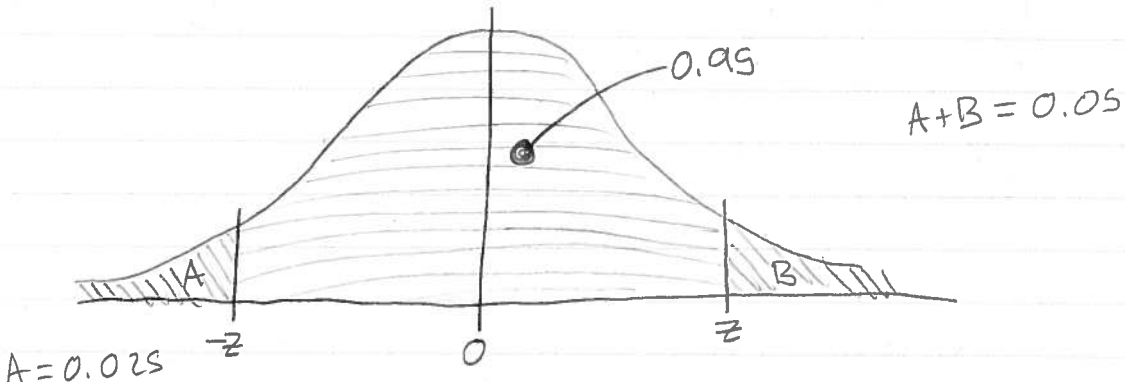
L_1 and L_2 have to be computable from the data. In this case we say $[L_1, L_2]$ is a confidence interval for μ and its confidence level is 95%: 95% confidence interval.

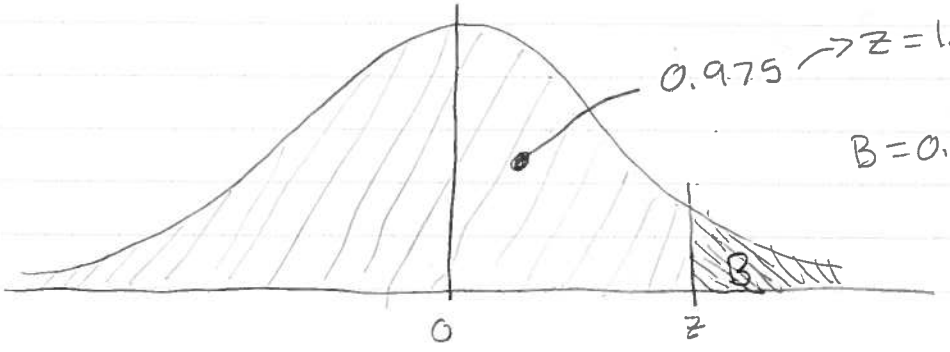
To build such an interval, we use the central limit theorem:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx Z \text{ (normal } (0,1))$$

In chapter 7, we looked at the following "inverse reading" problem (using Table 17.3). Find z such that

$$P(-z \leq Z \leq z) = 0.95$$





$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

Recall that: $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \approx Z$

$$P\left(-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right) \approx 0.95$$

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96$$
$$\bar{X}-\mu \leq 1.96(\sigma/\sqrt{n})$$
$$\mu \geq -(1.96(\sigma/\sqrt{n})) + \bar{X}$$

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \geq 1.96$$
$$\bar{X}-\mu \geq 1.96(\sigma/\sqrt{n})$$
$$\mu \leq -(1.96(\sigma/\sqrt{n})) + \bar{X}$$

$$P\Big(\underbrace{\bar{X}-1.96(\sigma/\sqrt{n})}_{L_1} \leq \mu \leq \underbrace{\bar{X}+1.96(\sigma/\sqrt{n})}_{L_2}\Big) \approx 0.95$$

This means that the 95% confidence interval for μ if σ^2 is known is:

$$\left[\bar{X}-1.96(\sigma/\sqrt{n});\ \bar{X}+1.96(\sigma/\sqrt{n})\right]$$

The interval is centered around $\bar{X}$. The value $\sigma/\sqrt{n}$ dictates how long the interval will be. Note that $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ is called the standard error of the mean.

Remark: since this interval was obtained using the central limit theorem, we need n to be large.

If n is small (ie. $n<30$), then we need to assume that X is normal.

Example 1 cont. assume that X is normal. Find the 95% confidence interval for μ. $\sigma = 115g$

General form: $\bar{x} \pm 1.96(\sigma/\sqrt{n})$

$= 798.8g \pm 1.96(115g/\sqrt{5})$

$= 798.8g \pm 100.8g$

$[698.0, 899.6]$

With 95% probability, the average cub weight is between 688.0g and 899.6g.

10.2 - Confidence Intervals for μ

σ^2 is unknown.

- we replace σ with its estimator S. In this case, the central limit theorem does not hold. But something else holds: $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a Student T distribution with $n-1$ degrees of freedom.

under the additional assumption that X is normal.

Therefore, the 95% confidence interval for μ is:

$$\bar{x} \pm t(S/\sqrt{n})$$

where t is read from Table 17.4 such that

$$P(-t \leq T_{n-1} \leq t) = 0.95$$

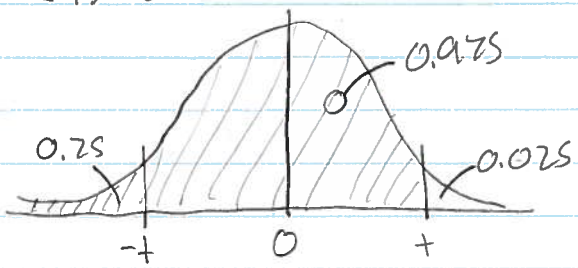
Table 17.4 gives the probability $P(T \leq t)$ for a random variable T with a Student T distribution of ∂ degrees of freedom, where $\partial = 1, 2, \ldots, 30$

	0.6	0.75	0.9	0.95	0.975	0.99	0.995
k				$t_{0.5,10}$			
1	0.325						
2							
3							
⋮							
10				1.1812			
⋮							
30							

$\begin{cases} T \text{ has 1 degree of freedom} \\ P(T \leq 0.325) = 0.6 \end{cases}$

$\begin{cases} T \text{ has 10 degrees of freedom} \\ P(T \leq 1.1812) = 0.95 \end{cases}$

For example, take a sample of size $n=20$. Then we look on row $n-1 = 19$. For the 95% c.i. we need to find t such that $P(-t \leq T \leq t) = 0.95$. This means that $P(T \leq t) = 0.975$



row: 19
column: 0.975
$\Big\} t = \cancel{1.729} \; 2.093$

MAT2379 11/11/2013

Confidence Interval for μ (§ 10.2)

The 95% confidence interval for μ is given by:

$$\bar{x} \pm t\,(s/\sqrt{n})$$

$\bar{x}$ = sample mean
s = sample std. dev.
n = sample size

t is read in table 17.4 such that:

$$P(-t \le T_{n-1} \le t) = 0.95$$

T_{n-1} is a random variable with T distribution with $n-1$ degrees of freedom

Therefore $P(T_{n-1} \le t) = 0.975$

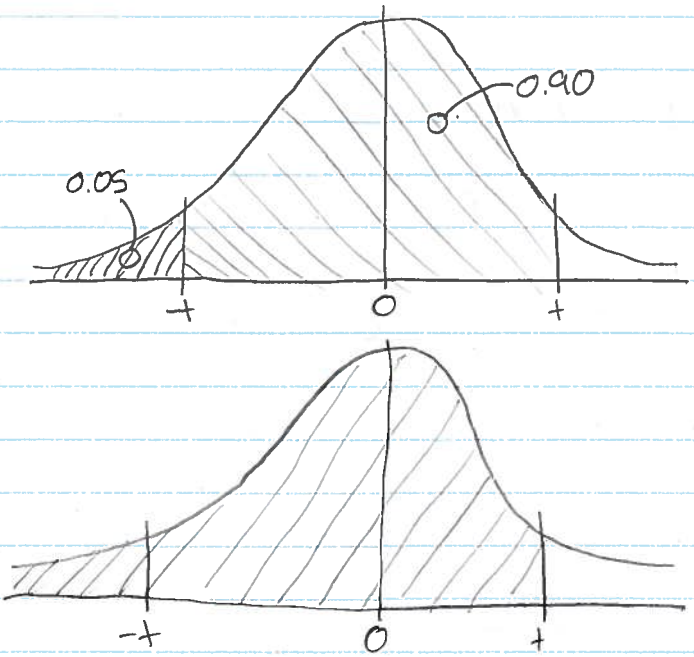
Example X = amount of butter fat (in lbs). Sample of $n = 20$ measurements from X (see notes for values). Find a 90% confidence interval for the average amount μ of butterfat (for the entire pop. of cows). The summary of the data:

$n = 20$
$\bar{x} = 507.5$ lbs
$s^2 = 8055.0625$ lbs^2

$$\bar{x} \pm t\,(s/\sqrt{n})$$
$$= 507.5 \pm \_\_\_\_\_\,(\sqrt{8055.0625}/\sqrt{20})$$

we have to find t such that

$$P(-t \le T_{19} \le t) = 0.90$$



$$P(T_{14} \leq +) = 0.95$$

Table 17.4:

row: 19
column: 0.95 } $+ = 1.729$

$$\bar{x} \pm +(s/\sqrt{n})$$
$$= 507.5 \pm 1.729\left(\frac{\sqrt{80055.0625}}{\sqrt{20}}\right) = [472.8, 542.2]$$

Confidence Intervals for the Proportion (§ 10.3)

p = proportion of individuals with a certain characteristic
draw a sample of size n
Y = total number of individuals (in the sample) who have this characteristic
An estimator for p is: $\hat{p} = Y/n$

ex
p = obese children in Canada aged 2-13
n = 975 children
y = 78 obese
$\Rightarrow \hat{p} = \frac{78}{975} = 0.08 = 8\%$

A 95% confidence interval for p is:

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $P(-z \leq Z \leq z) = 0.95$, z is found in table 17.3

ex Find a 90% confidence interval for p. We need to find z such that $P(-z \leq Z \leq z) = 0.90$
Hence $P(Z \leq z) = 0.95$
We find $z = 1.645$

$$c.i. = \hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= (0.08) \pm (1.645)\sqrt{(0.08)(1-0.08)/(975)}$$
$$= [0.0657, 0.0943)$$
$$= [6.57\%, 9.43\%)$$

∴ with probability of 90%, we can say that the proportion p of obese children is between 6.57% and 9.43%.

Hypothesis Testing - Ch. 11

We confront some hypotheses H_0 and H_1 which speak about the values of an unknown parameter in the population (eg. μ or proportion p).

We formulate H_0 and H_1 with hope of being able to reject H_0 (and therefore gain evidense for H_1)

H_0 is called the null hypothesis
H_1 is called the research hypothesis.

Example A new drug which is produced to reduce the systolic blood pressure in a certain population below the value 130,

X = systolic blood pressure of a randomly chosen patient who is using this drug.

$\mu = E(X)$ is the average systolic blood pressure of all patients using this drug (ie. the entire pop.)

$H_0: \mu = 130$ ↖ we want to reject this

$H_1: \mu < 130$ ↖ we want to gain evidence for

We have 4 situations:

	H_0 is true	H_1 is true
Reject H_0	Type I Error (α)	Correct Decision
Fail to reject H_0 (accept)	Correct Decision	Type II Error (β)

Ex 1: $H_0: \mu = 130$ $H_1: \mu < 130$

Type I error occurs when we decide that the new drug has lowered the blood pressure (ie. reject H_0) when in fact it has not.

Type II error occurs when we fail to gain evidence that the new drug is efficient, when in fact it is.

To develop the theory, we consider 2 cases:

1) σ^2 is known (section 11.1) → skip b/c it never really happens.
2) σ^2 is unknown (section 11.2)

Hypothesis Testing for μ when σ^2 is Unknown (§ 11.2)

We will consider H_0 of the form: $H_0: \mu = \mu_0$, where μ_0 is a known value.

We have 3 cases for H_1:

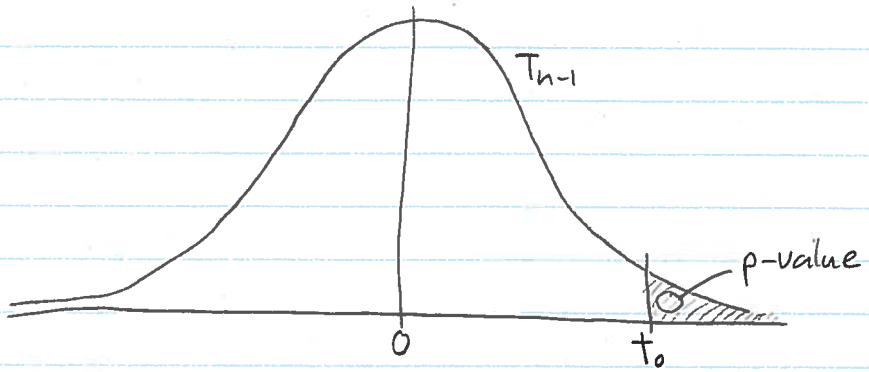
Case I: $H_0: \mu = \mu_0$ $\quad H_1: \mu > \mu_0$
Case II: $H_0: \mu = \mu_0$ $\quad H_1: \mu < \mu_0$
Case III: $H_0: \mu = \mu_0$ $\quad H_1: \mu \neq \mu_0$

Case I: $H_0: \mu = \mu_0$ $\quad H_1: \mu > \mu_0$

we look at a sample of size n,
we compute for this sample $\bar{x}$ and s
compare $\bar{x}$ (which is an approximation of μ) with μ_0
If $\bar{x} < \mu_0$, then we do not have evidence for H_1 Stop!
If $\bar{x} > \mu_0$, then we compute

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

A "large" value of t_0 means strong evidence for H_1
This t_0 should be compared with all the other values that could be obtained when we are repeating this procedure. These other values follow a T_{n-1} dist (§ 10.2)



To be able to say when t_0 is large, we define:

p-value $= P(T_{n-1} > t_0)$

t_0 is large (and hence we have evidence for H_1) : if p-value is small, which usually means p-value $< 0.05 = \alpha$.

0.05 is called the significance level of the test.

The rule is : if p-value $< \alpha$, then reject H_0 (in favor of H_1)
if p-value $> \alpha$, then we fail to reject H_0, we do not have enough evidence for H_1.

α is given
this rule ensures that the probability of Type I error is α.

comment: p-value is a great tool (even without comparing it with any α); the smaller the p-value, the stronger evidence we have for H_1

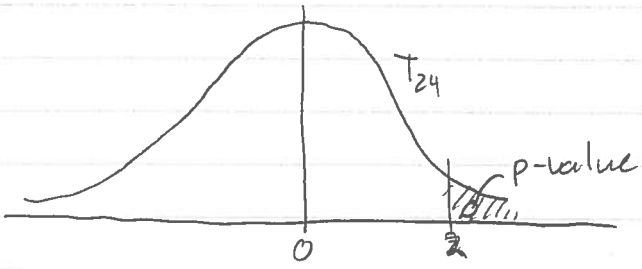
This is called a right-tail test.

Example:
- right-tail test
- X = benzene exposure level
- the standard requires that the level of benzene does not exceed the value of 1 ppm
- n = 25 workers (oil refining industry)
- $\bar{x} = 1.03$ ppm
- $s = 0.075$ ppm
- Is there enough evidence that in this refinery where these people are working, the average level of exposure is higher than 1 ppm? use $\alpha = 0.05$

$H_0 : \mu = 1 \quad H_1 : \mu > 1$

Since $1.03 > 1.0$, we can proceed with statistics

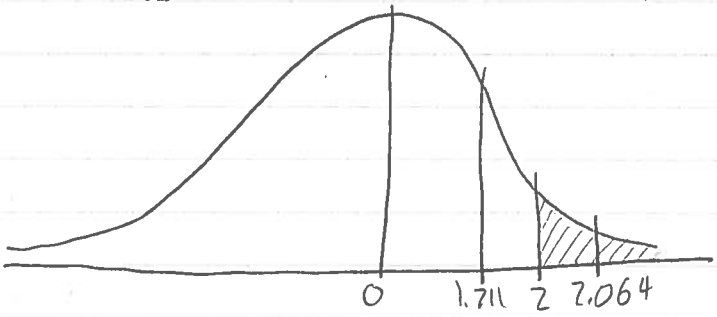
$$t_0 = \frac{1.03 - 1.00}{0.075/\sqrt{25}} = 2.00$$



p-value $= P(T_{24} > 2.00)$

look in table 17.4:

2.00 falls b/t 1.711 (area to the right is 0.05) and 2.064 (area to the right is 0.025)



$0.025 < \text{p-value} < 0.05$

since p-value $< 0.05 = \alpha$, we reject H_0. We have evidence that $\mu > 1$

MAT2379 13/11/2013

Hypothesis Testing for μ when σ^2 is Known (§11.2)

Case I: $H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$

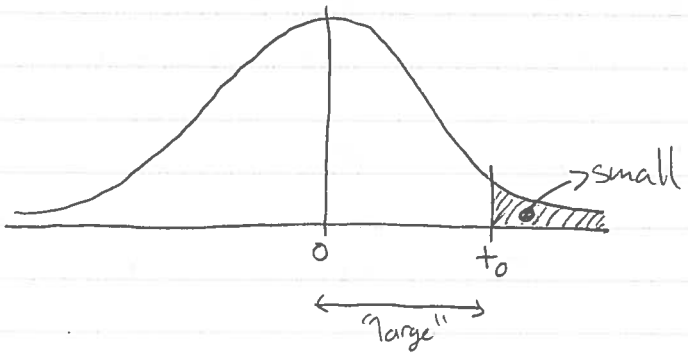
we compute $\frac{\bar{x}-\mu_0}{s/\sqrt{n}} = t_0$; the observed value of the test statistic T_0

here, the test statistic is: $T_0 = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$

if t_0 is large, then we reject H_0 in favour of H_1

p-value $= P(T > t_0)$

we say that t_0 is large if p-value is small



"small" means smaller than a given value α (eg. $\alpha = 0.05$).

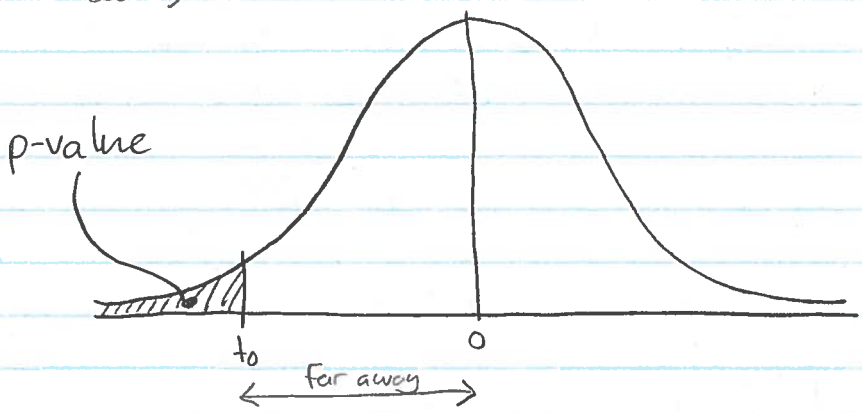
In summary,

- we reject H_0 if p-value $< \alpha$
- we fail to reject H_0 if p-value $> \alpha$

Case II: $H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$

we still compute $t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ if $\bar{x} < \mu_0$

in this case, $t_0 < 0$



we define: p-value = $P(T < t_0)$

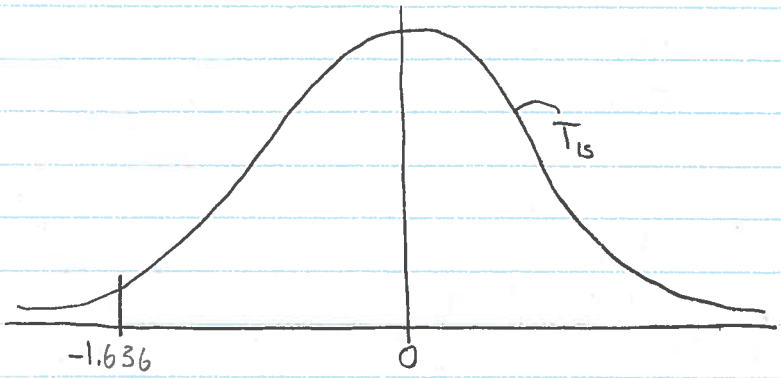
Rule: we reject H_0 if p-value $< \alpha$
we fail to reject H_0 if p-value $> \alpha$

This is called a left-tail test.

Example 1 A new drug is supposed to reduce the systolic blood pressure.
16 patients used this drug ✱cont. from last class.
$\bar{x} = 123.7$
$n = 16$
$s = 15.4$
Is there enough evidence that the new drug is efficient in reducing the systolic blood pressure. Use $\alpha = 0.01$

$H_0: \mu = 130$ $H_1: \mu < 130$

$$t_0 = \frac{123.7 - 130}{15.4/\sqrt{16}} = -1.636$$



p-value $= P(T_{15} < -1.636)$

p-value $= P(T_{15} < -1.636)$
$= P(T_{15} > 1.636)$

Two ways to answer this:
1) Table 17.4 will give a range for this prob. (not an exact value)
row 15 → 1.341 and 1.753, 1.636 is in b/t these values

$P(T_{15} < 1.341) = 0.90 \rightarrow P(T_{15} > 1.341) = 1 - 0.90 = 0.10$
$P(T_{15} < 1.753) = 0.95 \rightarrow P(T_{15} > 1.753) = 1 - 0.95 = 0.05$

Hence: $P(T_{15} > 1.636)$ is in b/t 0.05 and 0.10

2) Use R: see assignment 4

conclusion: $0.05 < \text{p-value} < 0.10$
recall that $\alpha = 0.01$

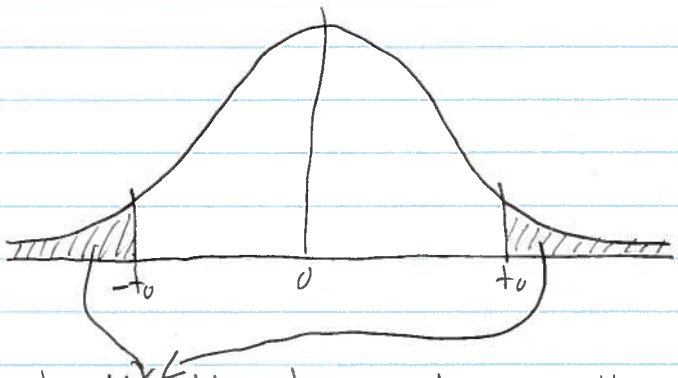
p-value $> \alpha$ → we fail to reject H_0. There is not enough evidence that the drug is efficient.

Case III: $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$

we compute $t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

$$\text{p-value} = 2P(T > |t_0|)$$
$$= \begin{cases} 2P(T > t_0) & \text{if } t_0 > 0 \ (\text{i.e. } \bar{x} > \mu_0) \\ 2P(T < t_0) & \text{if } t_0 < 0 \ (\text{i.e. } \bar{x} < \mu_0) \end{cases}$$

case when $t_0 > 0$:



together, these two prob. give the p-value

rule: $\begin{cases} \text{reject } H_0 \text{ if p-value} < \alpha \\ \text{fail to reject } H_0, \text{ otherwise} \end{cases}$

this is called a two-tailed test

Example 3: Measurements of blood viscosity were made on lab mice. A normal value should be around 3.95. A new drug is suspected to effect the blood viscosity level. Levels which are too small or too large (compared with 3.95) are not acceptable. A sample of $n=9$ mice is used. (Assume that the blood viscosity is normally distributed. We obtain: $\bar{x} = 4.25$, $s = 0.6$. Is there enough evidence that the average level of blood viscosity is significantly different compared to the value 3.95? Use $\alpha = 0.05$

$H_0: \mu = 3.95$ $\quad H_1: \mu \neq 3.95$

★ this is a two-tailed test

$$t_0 = \frac{4.25 - 3.95}{0.6/\sqrt{9}} = 1.5$$

$$p\text{-value} = 2P(T_8 > 1.5)$$

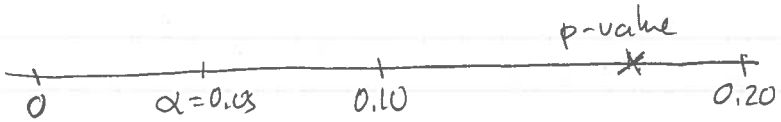
we first find the range for $P(T_8 > 1.5)$ using table 17.4:
in row 8, we find: $1.397 \rightarrow P(T_8 < 1.397) = 0.90$
$1.860 \rightarrow P(T_8 < 1.860) = 0.95$

$P(T_8 > 1.397) = 1 - 0.90 = 0.10$
$P(T_8 > 1.860) = 1 - 0.95 = 0.05$

The range for $P(T_8 > 0.15)$ is $(0.05; 0.10)$

The range for the p-value is $(0.10, 0.20) = 2(0.05, 0.10)$

Conclusion: $0.10 < p\text{-value} < 0.20$
Recall that $\alpha = 0.05$



p-value $> \alpha \Rightarrow$ we fail to reject H_0 and there is not enough evidence that the blood viscosity levels diviate significantly from the value 3.95, due to the new drug.

Hypothesis Testing for Proportions (§ 11.3)

p = proportion of individuals (in the population) who have a certain characteristic.

n = sample size

Y = number of people in the sample who have this characteristic.

$\hat{p} = \frac{Y}{n}$ = estimator for p; $z_0 = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$

Case I: $H_0: p = p_0$ $H_1: p > p_0$

$p\text{-value} = P(Z > z_0)$

Case II: $H_0: p = p_0$ $H_1: p < p_0$

$p\text{-value} = P(Z < z_0)$

Case III: $H_0: p = p_0$ $H_1: p \neq p_0$

$$p\text{-value} = \begin{cases} 2P(Z > z_0) & \text{if } z_0 > 0 \\ 2P(Z < z_0) & \text{if } z_0 < 0 \end{cases}$$

reject H_0 if $p\text{-value} < \alpha$

Example

Tarceva is an anti-cancer drug produced by Roche Holding.

Lung cancer: at most 10% will survive for 3 years after diagnostic.

150 patients → 22 survived for 3 years

$n = 150$

$Y = 22$

p = proportion of patients who survive for 3 yrs after using Tarceva

Is there enough evidence that the proportion p is higher than 10%? use $\alpha = 0.05$

$H_0: p = 0.10 \qquad H_1: p > 0.10$

$\hat{p} = 22/150 = 0.1467 > 0.10$

$$z_0 = \frac{0.1467 - 0.10}{\sqrt{\frac{0.10(1-0.10)}{150}}} = 1.91$$

MAT2379A 18/11/2013

R Code

```
>t.test(x)                                  => case III
>t.test(x, alternative="greater")           => case II
>t.test(x, alternative="less")              => case I
>t.test(x, mu=0.1)                          => for μ≠0
>t.test(x)$conf.int                         => just the confidence interval
>t.test(x, conf.lev=0.93)$conf.int
```

p-value $= P(Z > 1.91)$
$= 1 - P(Z < 1.91)$
$= 1 - (0.9719)$
$= 0.0291 < \alpha = 0.05$

$\therefore$ p-value $< \alpha$
$\therefore$ we reject H_0; there is enough evidence for H_1

MA7237A 18/11/2013

Office Hours on December 9th = 1:00 to 3:00

Notes on the formula sheet:

- statistic used for → confidence interval
  → tests

for mean μ of a normal popalation with known variance σ^2 is:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ has a standard normal dist.}$$

This means that

→ 95% confidence interval for μ is:

$$\bar{x} \pm z\left(\frac{\sigma}{\sqrt{n}}\right), \quad P(-z \leq Z \leq z) = 0.95$$

→ for tests, we decide what case we consider (ie. what is the alternative)

Case I $H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$

$$z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \qquad \text{p-value} = P(Z > z_0)$$

compare p-value with α

Chapter 12 - Comparison of Two Independent Samples

We want to compare the means μ_1 and μ_2 of two independent populations, using:

a) confidence intervals
b) tests of hypotheses

§12.2 - Confidence Intervals and Tests for Means

Population 1
- sample of size n_1
- sample mean $\bar{x}_1$
- sample variance s_1^2
- mean μ_1
- variance σ_1^2

Population 2
- sample of size n_2
- sample mean $\bar{x}_2$
- sample variance s_2^2
- mean μ_2
- variance σ_2^2

Goal: estimate $\mu_1 - \mu_2$ by $\bar{x}_1 - \bar{x}_2$

Interpretation: if $\bar{x}_1 - \bar{x}_2 > 0$, then we have some evidence that $\mu_1 > \mu_2$

Example 1 we want to see if there is a difference in the final exam grade between the male population (pop 1) and the female population (pop 2)

$n_1 = 37$ $\quad$ $n_2 = 30$

$\bar{x}_1 = 85.738$ $\quad$ $\bar{x}_2 = 89.4$

A point estimate for $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2 = -3.662 < 0$
We might infer that $\mu_1 - \mu_2 < 0$ but this is a very rough conclusion. We want to refine this.

We will use:

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = z$$

has approximately a standard normal distribution

We have four cases:

Case I - Populations are normal with known variances

c.i. for $\mu_1 - \mu_2$ is:

$$\bar{x}_1 - \bar{x}_2 \pm z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

★ this is not a realistic case ★

Case II - Populations are normal,
- unknown variances σ_1^2, σ_2^2 and
- $\sigma_1^2 = \sigma_2^2$

Case III - Populations are normal,
- unknown variances σ_1^2, σ_2^2, but
- $\sigma_1^2 \neq \sigma_2^2$

Case IV - Populations are arbitrary
- unknown variances σ_1^2, σ_2^2

MAT2379A 20/11/2013

Chapter 12.2: Confidence Intervals and Tests for Means

Two independent populations with means μ_1, μ_2 and variances σ_1^2, σ_2^2

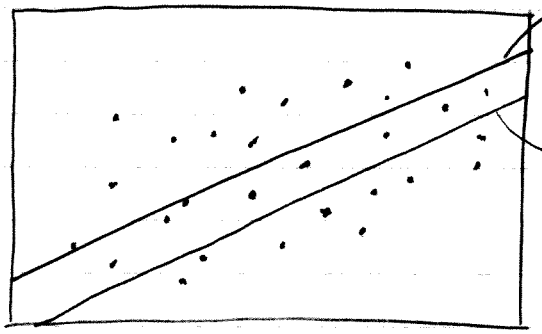
Two Samples:

Sample 1
- sample size n_1
- sample mean $\bar{X}_1$
- sample variance S_1^2

Sample 2
- sample size n_2
- sample mean $\bar{X}_2$
- sample variance S_2^2

Case II

- inference on $\mu_1 - \mu_2$: the two populations are normal with equal variances. (ie. $\sigma_1^2 = \sigma_2^2 = \sigma^2$)
- to check the validity of this assumption, we use overlaged QQ plots (with R)



→ QQ plot and the line of best fit for the 1st sample

→ QQ plot and the line of best fit for the 2nd sample.

- we need
  - → both plots to be linear
  - → the lines to be parallel for $\sigma_1^2 = \sigma_2^2$

A) Interval Estimation for $\mu_1 - \mu_2$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2(1/n_1 + 1/n_2)}}$$ has approx. a standard normal distribution.

★ σ^2 is unknown! ★

σ^2 is replaced by the following estimator:

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

(S_p^2: p = pooled)

S_p^2 is called the pooled variance and is used only when the assumption of equality of variances is verified.

$$T_0 = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(1/n_1 + 1/n_2)}}$$ has a T-distribution with n_1+n_2-2 degrees of freedom.

not on formula sheet

Hence, the 95% confidence interval for $\mu_1 - \mu_2$ is:

$(\bar{X}_1 - \bar{X}_2) \pm t\sqrt{S_p^2(1/n_1 + 1/n_2)}$, where t is found in Table 17.4 such that $P(-t \leq T_{n_1+n_2-2} \leq t) = 0.95$

Interpretation of the interval:
- → if the interval contains only positive values, then we can say that $\mu_1 > \mu_2$ with prob. 95%
- → if the interval contains only negative values, then we can say that $\mu_1 < \mu_2$ with prob. 95%
- → if the interval contains $\emptyset$, then we can not draw any conclusions.

Example 1

X_1 = the score on a math test taken by a randomly chosen student in a large high school.

X_2 = the score on a math test taken by a randomly chosen student in a small high school

μ_1 = mean of X_1

μ_2 = mean of X_2

σ_1^2 = var (X_1)

σ_2^2 = var (X_2)

Assume that the two populations X_1 and X_2 are normal with equal variance. Hence, we are in case II.

$n_1 = 9$ $\quad n_2 = 15$

$\bar{x}_1 = 81.31$ $\quad \bar{x}_2 = 78.61$ $\quad$ note: $\bar{x}_1 > \bar{x}_2$

$s_1^2 = 60.76$ $\quad s_2^2 = 48.24$

Compute a 95% confidence interval for $\mu_1 - \mu_2$. Interpret the result.

$$(\bar{x}_1 - \bar{x}_2) \pm t \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

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

$$= \frac{(9-1)(60.76) + (15-1)(48.24)}{9 + 15 - 2}$$

$$= 52.79$$

We look for t with Table 17.4 or with R

→ row: $9 + 15 - 2 = 22$

→ column: 0.975

$\Big\}$ $t = 2.074$

$$(\bar{x}_1 - \bar{x}_2) \pm t \sqrt{s_p^2 (1/n_1 + 1/n_2)}$$

$$= (81.31 - 78.61) \pm (2.074) \sqrt{52.79 (1/9 + 1/15)}$$

$$= [-3.65 ; 9.05]$$

Since the interval contains 0, we cannot say that μ_1 is greater or smaller than μ_2

b) Hypothesis Testing

$$T_0 = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(1/n_1 + 1/n_2)}}$$ has a T-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Case I: $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$ (right tail test)

we compute

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{s_p^2(1/n_1 + 1/n_2)}}$$

- a large value of the difference $(\bar{x}_1 - \bar{x}_2) - 0$ is an indication that H_1 may be true.
- how large should the value be to reject H_0?

p-value $= P(T_{n_1+n_2-2} > t_0)$

if p-value $< \alpha$, then we reject H_0

if p-value $> \alpha$, then we fail to reject H_0

Case II: $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$ (left tail test)

p-value $= P(T_{n_1+n_2-2} < t_0)$

Case III: $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$ (two tailed test)

p-value $= \begin{cases} 2P(T > t_0) & \text{if } t_0 > 0 \\ 2P(T < t_0) & \text{if } t_0 < 0 \end{cases}$

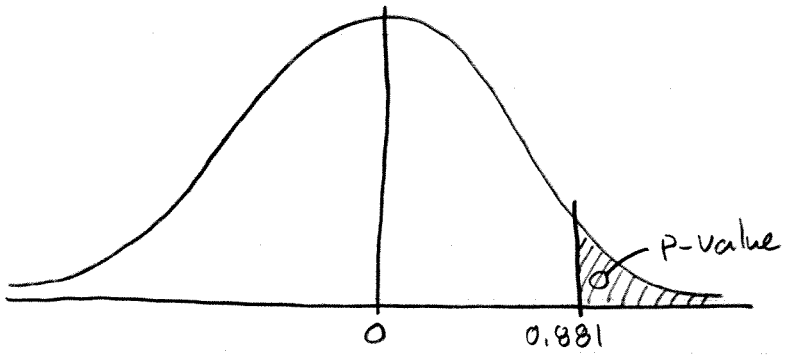
Example 1 Cont.

Compare the means μ_1 and μ_2 from the point of view of hypothesis testing. Is there enough evidence that μ_1 is greater than μ_2? Use $\alpha = 0.05$

$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$

$s_p^2 = 52.79$

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_0) - 0}{\sqrt{s_p^2 (1/n_1 + 1/n_2)}} = \frac{81.31 - 78.61}{\sqrt{52.79 (1/9 + 1/15)}} = 0.881$$

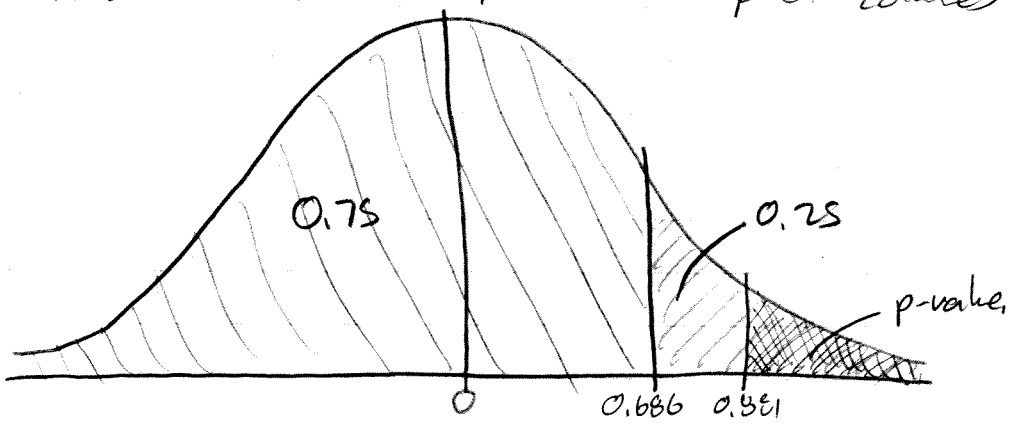


p-value $= P(T_{22} > 0.881)$ → look in table 17.4

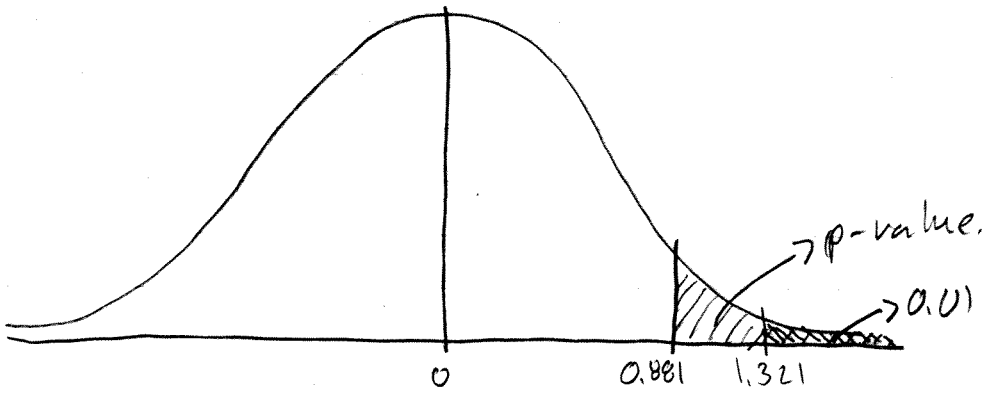
row: 22

find: 0.686 → 0.75 (on top of column)

find: 1.321 → 0.9 (on top of column)

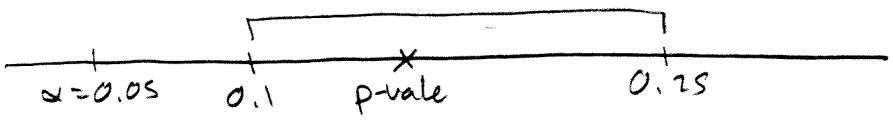


p-value < 0.25



p-value > 0.1

$0.1 <$ p-value < 0.25



p-value $> \alpha$, so we fail to reject H_0
we don't have enough evidence that $\mu_1 > \mu_2$
at level $\alpha = 0.05$

Case III: Normal Populations with unequal variances

- populations are still normal (ie. the two QQ plots seem to be linear) but the variances are not equal (ie. the 2 lines do NOT seem to be parallel).

a) Confidence Intervals

$$T_0 = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

has a T distribution with an ugly number of degrees of freedom ∂ [nu:]

$$\partial = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{S_2^2}{n_2}\right)^2}$$

MAT2379A 25/11/13

cont...

confidence interval for $\mu_1 - \mu_2$ is: $\bar{x}_1 - \bar{x}_2 \pm t\sqrt{s_1^2/n_1 + s_2^2/n_2}$

t is found in table 17.4
row is ∂ (round down)
column is 0.975 for a 95% confidence interval

Example 3

X_1 = blood volume (in ml) for a paraplegic men who participates in vigurous physical activities.
X_2 = blood volume (in ml) for an able-bodied man who participates in normal physical activities.

$n_1 = 7$ $n_2 = 10$ $\mu_1 = E(X_1)$
$\bar{x}_1 = 1511.714$ $\bar{x}_2 = 1118.4$ $\mu_2 = E(X_2)$
$s_1^2 = 49669.905$ $s_2^2 = 15297.6$

Is there evidence that $\mu_1 > \mu_2$? Use a 95% confidence interval to answer this.

Variance σ_1^2 and σ_2^2 are not equal

$$v = \frac{\left(\frac{49669.905}{7} + \frac{15297.6}{10}\right)^2}{\frac{1}{6}\left(\frac{49669.905}{7}\right)^2 + \frac{1}{9}\left(\frac{15297.6}{10}\right)^2} = 8.599 \longrightarrow 8 \text{ degrees of freedom}$$

$$\text{confidence interval} = \bar{x}_1 - \bar{x}_2 \pm t\sqrt{s_1^2/n_1 + s_2^2/n_2}$$
$$= (1511.714) - (1118.4) \pm t\sqrt{\frac{49669.905}{7} + \frac{15297.6}{10}}$$

t is read in Table 17.4
{ row 8
{ column 0.975 $t = 2.306$

95% confidence interval: [179.148, 607.480]

- $\mu_1 - \mu_2 > 0$ with prob. 95% since the interval contains only +ve values
- $\mu_1 > \mu_2$

Hypothesis Testing for $\mu_1 - \mu_2$ with $\sigma_1^2 \neq \sigma_2^2$

$T_0 = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$ has a $T(\partial)$ distribution (∂ is the same as before)

Case I: $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$

compute $t_0 = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$

p-value $= P(T_\partial > t_0)$

if p-value $< \alpha$, reject H_0

Case II: $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$

t_0 is the same as in Case I (but $t_0 < 0$)

p-value $= P(T_\partial < t_0)$

Case III: $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$

t_0 is the same as in Case I

p-value $= \begin{cases} 2P(T_0 > t_0) & \text{if } t_0 > 0 \\ 2P(T_0 < t_0) & \text{if } t_0 < 0 \end{cases}$

Ex 3 cont...

Justify the claim $\mu_1 > \mu_2$ using the method of hypothesis testing. Use $\alpha = 0.05$

$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$

$t_0 = \dfrac{(1511.714) - (1118.4)}{\sqrt{\dfrac{446611.905}{7} + \dfrac{152917.6}{10}}} = 4.234$

p-value $= P(T_8 > 4.234) < 0.005$

note $4.234 >$ the last value in T17.4 on row 8

Hence p-value < 0.05
We reject H_0. We have enough evidence for $\mu_1 > \mu_2$

Case IV:
- arbitrary populations
- unknown variances
- we skip this case.

Chapter 13 - Paired Samples

We study two dependent populations.
Typical examples of measurements X and Y which are dependent are made on the same subject.
These observations will come in pairs:

for subject 1: (x_1, y_1)
for subject 2: (x_2, y_2)
⋮
for subject n: (x_n, y_n)

examples:

{ x = weight before a weight-loss program
{ y = weight of the same person after the program
{ x = midterm grade
{ y = final exam grade for the same person
{ x = blood pressure before exercising
{ y = ——— " ——— after ——— " ———

Note: both X and Y samples have the same sample size.

We denote by $X_1, X_2, \ldots X_n$ the sample of X
We denote by $Y_1, Y_2, \ldots Y_n$ the sample of Y

We compute the differences:

$$D_1 = X_1 - Y_1$$
$$D_2 = X_2 - Y_2$$
$$\vdots$$
$$D_n = X_n - Y_2$$

We assume that these differences are normally distributed (check with a QQ Plot)

We want to compare μ_x with μ_y, where
$\mu_x = E(X)$ and $\mu_Y = E(Y)$
from the point of view of:
a) confidence intervals for $\mu_x - \mu_y = \mu_D$
b) hypothesis testing on μ_D

Conclusions:
a) If the interval contains only +ve value, then $\mu_D > 0$ with a high probability. This means that $\mu_x > \mu_Y$. Similar for -ve values. If the confidence interval contains both +ve and -ve values, we can't make any conclusions.
b) Test for:

or $H_0: \mu_D = 0$ against $H_1: \mu_D > 0$ $(\mu_x > \mu_y)$
or $H_0: \mu_D = 0$ against $H_1: \mu_D < 0$ $(\mu_x < \mu_y)$
$H_0: \mu_D = 0$ against $H_1: \mu_D \neq 0$ $(\mu_x \neq \mu_y)$

The conclusions are based on the SINGLE sample $d_1, d_2, \ldots, d_n$ of differences $d_i = x_i - y_i$

13.1 - Confidence Intervals for μ_D

From Ch. 10, the c.i. for μ_D is;

$$\bar{d} \pm t\left(s_d/\sqrt{n}\right)$$

where

$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$ → sample mean of differences

$s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$ → sample variance of the differences

t is read in Table 17.4: row $n-1$

Example 1

mcPP is thought to affect apetite and food intake.
$n = 9$ women

Session 1 = they took mcPP for 2 weeks, they recorded x = weight loss (x can be -ve: weight gain)

They had a break for 2 weeks.

Session 2 = the same people took a placebo for 2 weeks and they recorded their weight loss as y.

Data:

session 1 (x)	session 2 (y)	difference (d)
1.1	0	1.1
1.3	-0.3	1.6
1.0	0.6	0.4
1.7	0.3	1.4
1.4	-0.7	2.1
0.1	-0.2	0.3
0.5	0.6	-0.1
1.6	0.9	0.7
-0.5	0.2	0.7

$\bar{d} = (1/9)(1.1 + \ldots + (-0.7)) = 0.76$

$s_d^2 = \frac{1}{8}\left[\sum_{i=1}^{9} d_i^2 - 9\bar{d}^2\right] = \frac{1}{8}\left((1.1)^2 + \ldots + (-0.7)^2 - 9(0.76)^2\right)$

$s_d = 0.88$

95% confidence interval

$0.76 \pm t\left(0.88/\sqrt{9}\right)$

to find t, look in T17.4

row 8

column: 0.975 $\Big\} t = 2.306$

$0.76 \pm 2.306\left(0.88/\sqrt{9}\right)$

$= [0.08, 1.44]$

∴ we are 95% confident that $\mu_D = \mu_x - \mu_y$ is in the interval $[0.08, 1.44]$

∴ this interval contains only +ve values we can say, with 95% confidence, that $\mu_D > 0$ i.e. $\mu_x > \mu_y$

∴ mcPP is efficient

13.2: Hypothesis Testing for μ_D

Case I: $H_0: \mu_D = 0$ $(\mu_x = \mu_y)$ $\quad H_1: \mu_D > 0$ $(\mu_x > \mu_y)$

$T_0 = \frac{\bar{D} - \emptyset}{S_d/\sqrt{n}}$ $\Big\}$ use theory from ch. 11 with $\mu_0 = \emptyset$

We compute the observed value:

$t_0 = \frac{\bar{d}}{S_d/\sqrt{n}}$ $\quad$ p-value $= P(T_{n-1} > t_0)$

if p-value $< \alpha$, reject H_0

Case II and III are similar (see posted notes)

Example 2

$n = 30$ diabetics

x = blood glucose level before they were taught how to use the meter

y = ~~"}~~ after ~~"}~~

$\bar{d} = 2.78$

$s_d = 6.05$

Is there sufficient evidence that the monitor helps reduce blood glucose levels, Use $\alpha = 0.01$

$H_0: \mu_D = 0$ $\quad H_1: \mu_D > 0$ $\quad (\mu_x - \mu_y)$

$$t_0 = \frac{2.78}{6.05/\sqrt{30}} = 2.52$$

p-value $= P(T_{29} > 2.52)$ is b/t 0.01 and 0.005

$\therefore$ p-value $< \alpha$

$\therefore$ we reject H_0

see summary on webpage of MAT2379 B

Chapter 14 - Categorical Data & Contingency Tables 27/11/2013

Data is classified according to 2 categorical variables,
Data is displayed in a $r \times c$ contingency table where
r = number of levels of the first variable
c = " " second "

Example 1

$n = 130$ individuals in a small Texas community
X = income level has 2 levels
- low income
- high income

Y = attitude regarding stricter gun regulation has 2 levels
- approve
- disapprove

	approve	disapprove	total
low income	18	48	66 (rand.)
high income	42	22	64 (rand.)
total	60 (rand.)	70 (rand.)	130

Test of Independence (14.1)
H_0: X and Y are independent
H_1: there is an association between X and Y

Example 2

39405 women between the ages of 50 and 59
They were assigned to two groups
19711 had exam + mamogram
19694 had exam only

	death	no death	total
exam + mammogram	107	19604	19711 (fixed)
exam only	105	19589	19694 (fixed)
total	212 (random)	39102 (random)	39405

This is a so called test of homegeneity (14.2)

Chapter 14.1 – Test of Independence

H_0: X and Y are independent
H_1: there is an association between X and Y

Notation: $n_{i\cdot} = \sum_{j=1}^{c} n_{ij}$ is the total for row i

n_{ij} = number of observations that fall in the cell (i,j) of the table

$n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$ is the total for column j

of course $n = \sum_{i=1}^{r} n_{i\cdot} = \sum_{j=1}^{c} n_{\cdot j}$ n = total sample size

We say that events A and B are independent if
$P(A \cap B) = P(A)P(B)$

We denote by:
p_{ij} = prob. that an obs. falls in cell (i,j)
$p_{i\cdot}$ = prob. that an obs. falls on row i
$p_{\cdot j}$ = prob. that an obs. falls on column j

H_0: $p_{ij} = p_{i\cdot} p_{\cdot j}$ for all i,j

Estimator of $p_{i\cdot}$ is $\hat{p}_{i\cdot} = n_{i\cdot}/n$

Estimater of $p_{\cdot j}$ is $\hat{p}_{\cdot j} = n_{\cdot j}/n$

If H_0 is true, then
$\hat{p}_{ij} = \hat{p}_{i\cdot}\,\hat{p}_{\cdot j}$ is an estimator for p_{ij}
and the expected number of obs. in cell (i,j) is

$$\hat{E}_{ij} = n\hat{p}_{ij} = n\hat{p}_{i\cdot}\hat{p}_{\cdot j} = n\left(\frac{n_{i\cdot}}{n}\right)\left(\frac{n_{\cdot j}}{n}\right) = \frac{n_{i\cdot}\,n_{\cdot j}}{n}$$

Note that $\hat{E}_{ij}$ is $\neq$ to the actual obs. number of obs. in cell (i,j)

$\hat{E}_{ij}$ = what you would expect to observe in cell (i,j), if X and Y are independent

To decide if the difference b/t the theoretical $\hat{E}_{ij}$ and the actual n_{ij} is significant, we compute:

$$u_o = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$ } observed value of the test statistic

u_o has a χ^2 [kai] square distribution with $(r-1)(c-1)$ degrees of freedom

p-value $= P(u_o > u_o)$

p-value $< \alpha \longrightarrow$ reject H_0

Table 17.5 gives the probabilities associated with χ^2 distribution with degrees of freedom $1, 2, 3, \ldots, 35$

The table gives $P(\chi^2 \leq u)$

Example 1 Cont....

Is there an association between the income level and the attitude towards handgun regulations? Use $\alpha = 0.05$

	approve	disapprove	total
low income	18 (30.46)	48 (35.34)	66
high income	42 (29.54)	22 (34.46)	64
total	60	70	130

The expected values under independence.

$$\hat{E}_{11} = \frac{u_{1.} n_{.1}}{n} = \frac{66 \times 60}{130} = 30.46$$

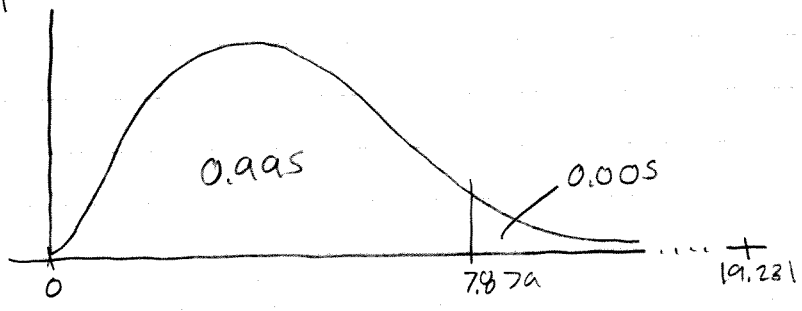
$$\hat{E}_{12} = \frac{n_{1\cdot} n_{\cdot 2}}{n} = \frac{66 \times 70}{130} = 35,34$$

$$\hat{E}_{21} = \frac{n_{2\cdot} n_{\cdot 1}}{n} = \frac{60 \times 64}{130} = 29,54$$

$$\hat{E}_{22} = \frac{n_{2\cdot} n_{\cdot 2}}{n} = \frac{70 \cdot 64}{130} = 34,46$$

$$u_0 = \frac{(18-34,46)^2}{34,46} + \frac{(48-35,34)^2}{35,34} + \frac{(42-29,54)^2}{29,54} + \frac{(22-34,46)^2}{34,46} = 19,231$$

p-value $= P(\chi^2(1) > 19,231) =$



p-value $<<< \alpha$ → reject H_0