COMP114 Experimental Methods in Computing

Experiment Design – Interpreting Results

Background

- Suppose we are interested in testing an hypothesis about “how good” a method, P, is at carrying out a given task.
- Assuming a precise definition of what is intended by “how good” has been given, the hypothesis is examined by running a series of experiments using random data.
- The results obtained are claimed to support the hypothesis.

Question –

- How can this claim be justified?
- Possible objections –

a. Random data was used and the results are just a coincidence.

b. The random data was “biased”.

Possible answers

- To answer objection (a) one approach is to reason that the “chance” of the results being coincidence is “too small”.
- For example –

  If a die is thrown 6000 times one would expect each of the 6 possible outcomes to occur “roughly” 1000 times.

  If, however, a 6 was thrown on 5000 of the tests, then “most people” would agree that the die used was “biased”.

Problematic issue –

- 5000 identical outcomes when only 1000 had been "expected" is accepted as indicating "bias" in the die used.
- But, what if the the results had been –

  3000 identical *OR* 2000 identical *OR*

  1500 *OR* 1250 *OR* 1100 … ?
- In other words,

  How large must the discrepancy between what is expected to happen (*1000 ´ 6s*) and what happens in practice *(???? ´ 6s*) be in order to be "confident" that the die is biased?

Deciding between "coincidence" and "real behaviour"

- In statistics the notion of "statistical significance" has been developed as a method of answering this question.
- We will describe some of the standard ideas used to assess experimental results by statistical methods.
- Our interest is in *applying* these ideas when conducting an experiment.
- We do not consider "*mathematical*" properties

Statistical approach – overview I

- Examining a claim such as:

  “The *outcome* of this experiment will be x”
- For example,

➢ “*If I add up the values that appear when this die is thrown 100 times then the total will be 350*”

- Or, in a more general form
- “*If I add up the values that appear when this die is thrown n times then the total will be 3.5×n*”
- The claim being studied is called the

  Null Hypothesis (*or N.H.*)

Statistical approach – overview II

- The experiment indicates:

  “The *result* of this experiment was y”
- For example,

➢ “*On adding the values that appeared when this die was thrown 100 times the total was 380*”

Statistical approach – overview III

- In total, we have –
- An *expected outcome* – x.
- An *actual outcome* – y.
- The *number of trials* made – n.
- Sometimes "expected" and "actual outcome" are stated as "averages" e.g. in die throwing case

x = 3.5

y = (Total thrown)/(Number of throws)

n = Number of times die was thrown

Statistical "significance" I

- The term "*significance*" refers to whether the difference between an *expected outcome* and the *actual outcome* is extreme enough to suggest that the *Null Hypothesis* is "incorrect".
- This is described by considering how "likely" such a discrepancy between the two would be.

Statistical “significance” II

- Conventionally three levels are used in practical experimental studies –

  Significant = =5% likelihood

  Highly significant = =1% likelihood

  Very highly significant = =0.1%
- In very informal terms, these express,
- “*That* x *is true given that* y *was observed is a 1-in-20/1-in-100/1-in-1000 chance*”

Summary I

- To decide if a hypothesis is acceptable or not (and the level these hold) by a series of experiments, we need:

A. The *expected result* – x (Null Hypothesis)

B. The *actual result* – y (experiment)

C. The *number of trials* used to find y (n)

D. The *probability* of y being the outcome after n trials whose predicted result was x.

Summary II

- The main complication that arises is computing the value from which conclusions will be drawn, i.e.

D. The *probability* of y being the outcome after n trials whose predicted result was x.

Some statistics jargon –

- In order properly to deal with the computational problem raised by finding

  “The *probability* of the outcome y after n trials whose predicted result was x.”

  we first need to introduce some terms from statistics.

Populations, Samples, Distributions

- *Population* – P

  The *collection* (set) of *possible outcomes*, e.g. the six possible results from throwing a die, the two possible ways a coin may land.
- *Sample* – x

  A member of a population, P.
- *Distribution* – D

  A *function* describing for each sample in the population, its *probability*.

The Uniform Distribution

- In the *Uniform Distribution* – each sample of the population is equally likely, e.g.

  P = {1,2,3,4,5,6}

  Prob[x chosen] = 1/6

  P={Heads,Tails}

  Prob[Heads]=Prob[Tails]=1/2
- The uniform distribution is often used in experiments dealing with estimating the “typical performance” of a program.

Some more jargon …

- In comparing the outcome (y) against the predicted outcome (x), some mechanism for considering how the values that led to y are "*spread*".
- The concept of "*spread*" is formally described by the ideas of *variance* and *standard deviation* (or *standard error*).

Average and Variance

- Given n samples from a population P - $X=<x_1,x_2,x_3,\ldots,x_n>$ - (assuming a uniform distribution)
- The *average* (or *mean*) value of X is

$$E(X) = (x_1+x_2+x_3+\ldots+x_n)/n$$

- The *variance*, V(X) of X is

$$[(x_1-E(X))^2+(x_2-E(X))^2+\ldots+(x_n-E(X))^2]/n$$

Average and Variance

- The notation E(X) arises from this value often being referred to as "*the expected value of a sample*".
- In the examples and cases looked at we will assume that the population is sampled using a *Uniform Distribution*.
- In a number of studies, however, the Uniform distribution over P is replaced by by one which "*biases*" to a subset of P.
- In such cases, a refinement of the definitions for E(X) and V(X) which considers this bias must be used.

Standard Deviation

- Variance provides one method for describing the "spread" of a range of samples drawn (uniformly) from a population.
- Standard deviation (which is closely related to variance) is another.
- Given n samples from a population P – $X=<x_1,x_2,x_3,\ldots,x_n>$ – the *standard deviation* of X is

$$S.D(X) = [V(X)]^{(1/2)}$$

- i.e. the *square root* of variance. The notation $\sigma(X)$ (or just σ) is sometimes used.

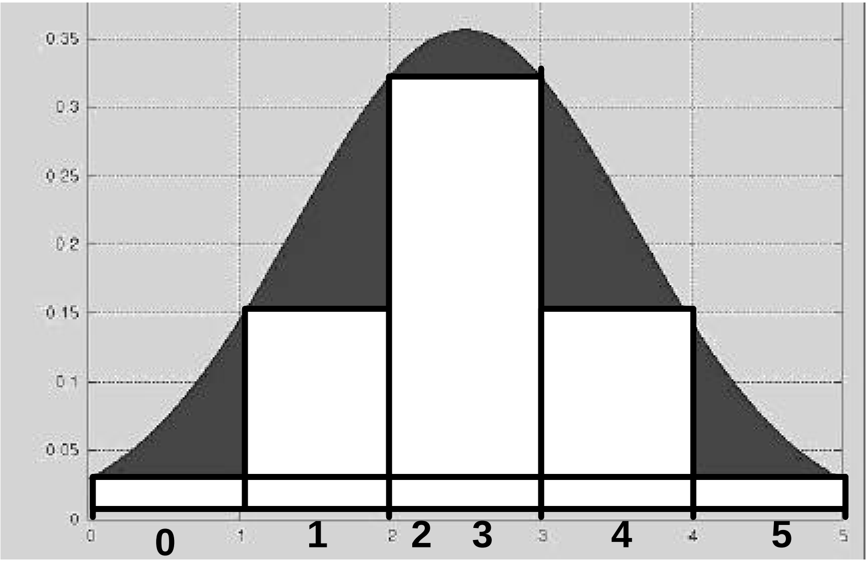
The Normal Distribution

- Suppose we are interested in how likely it is that some number of tests will produce the same result. For example,
- We have a single coin; in each test this may land Heads; if we carry out n tests how often might k Heads be seen?

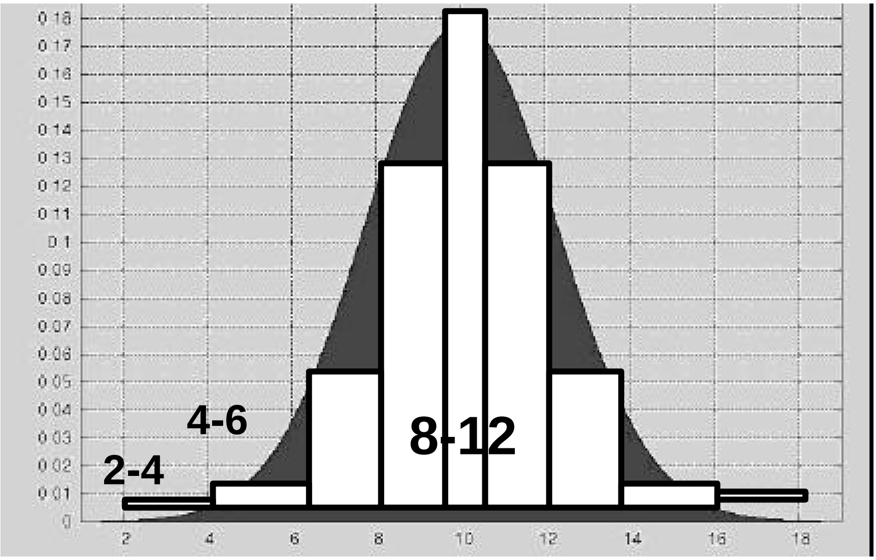
What are the possibilities with 5 tests?

- We can have 0, 1, 2, 3, 4 or 5 heads. If the coin is fair (equally likely to fall Heads or Tails), the chance of each is,

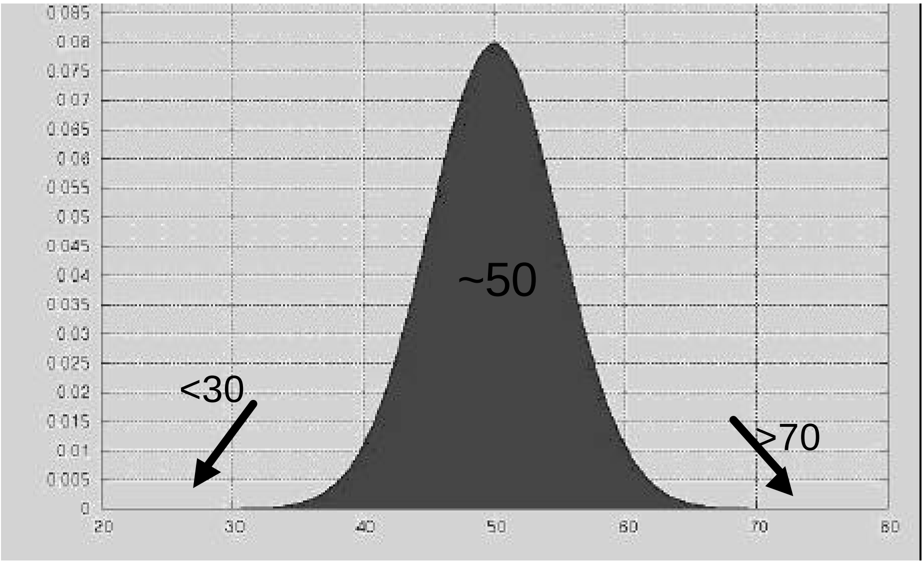
Number of Heads (5 *tests*)



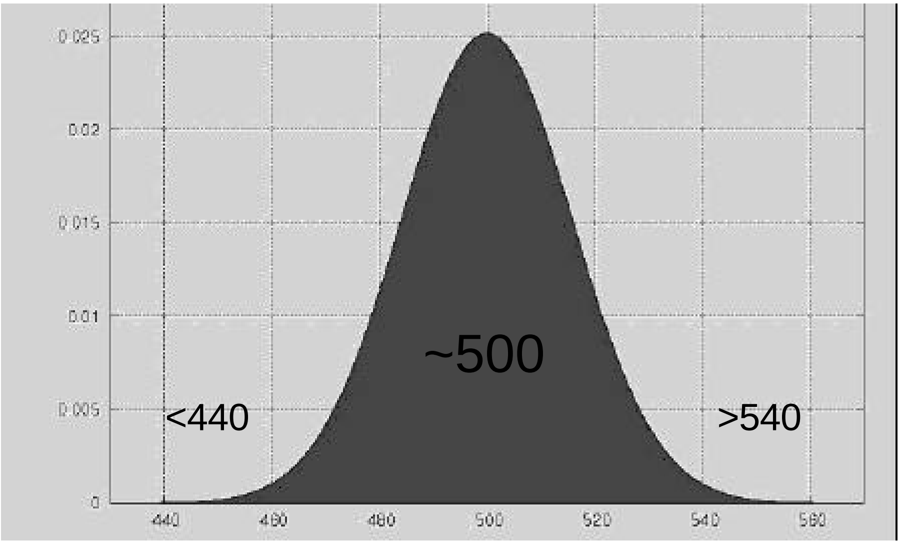
With 20 tests …



And 100 …



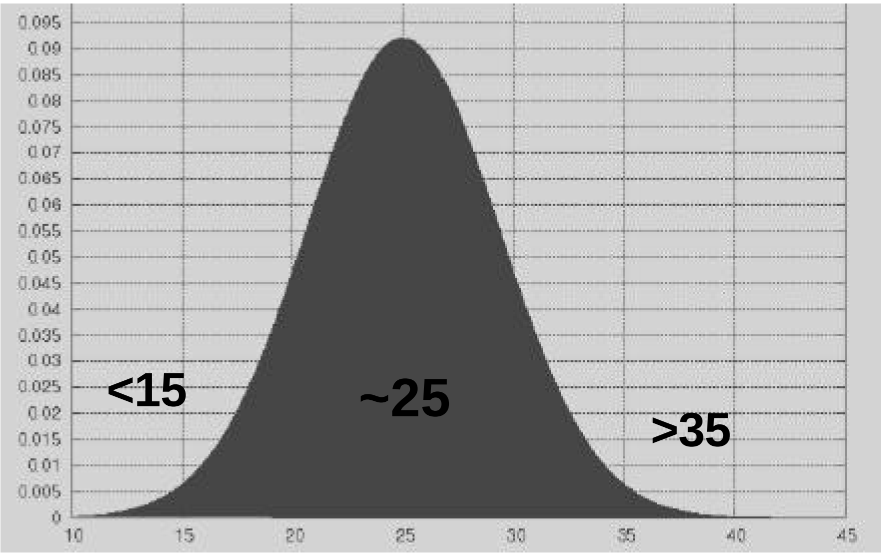
Finally, for 1000 tests …



Properties –

- The term “Normal distribution”, is used for the family of “bell-shaped” curves with a number of very important properties – e.g.
- As the number of tests (n) increases, the probability of the number of Heads seen being “noticeably” different from n/2 becomes very small.
- A similar behaviour would occur if a biased coin was used, so that if, for example, Heads occurred with probability 1/4 …

Biased coin, 100 tests



Properties of Normal Distribution

- The Normal Distribution is used/occurs in a number of practical situations – e.g.

  exam score ranges

- The actual “bell curve” formed is defined by two parameters –

  μ - The mean (average value)

  σ - The Standard Deviation

- The notations $N(\mu,\sigma)$ and $\Phi(x)$ are often used – the second being the value x maps to on the curve defined by $N(\mu,\sigma)$.
- The value $\Phi(x)$ defines the “chance of x occurring”.

Applying N(m, s) to Real Problems

- Recall that we are interested in dealing with the following problem –

  “Finding the *probability* of an outcome y after n tests with *predicted* result x.”

- How could assuming the behaviour of an experiment follows a Normal Distribution, help us?

Distance from the mean I

- From the results giving the chance of a coin landing Heads on *at least* k tests out of 100, we can compute the following Table.

k	Prob[≥k heads]
50	0.5
55	0.18165
60	0.03452
65	0.00319
70	0.00014
75	0.00001

Distance from the mean II

- When the experiment outcome (y – the number of Heads) is further and further from the predicted outcome (x), it strengthens the evidence that the Null Hypothesis (the coin lands Heads on x tests) is in fact *incorrect*.
- Is the increment “5” important in the Table on the previous slide?
- Yes. In the example, the mean is 50, and the standard deviation is ≈5

Distance from the mean III

- An outcome (60) at least 2 standard deviations away from the mean (50) is already “significant” (<1-in-20 chance)
- An outcome (65) at least 3 standard deviations away from the mean is already “highly significant” (<1-in-100 chance)
- An outcome (70) at least 4 standard deviations away is “very highly significant” (<1-in-1000 chance).
- So we could reject the claim that “the coin was fair” on the basis of an experiment giving such results.

Overview – basic approach.

- Given a population, P, whose mean *is claimed to be* μ and whose standard deviation σ *is known* the aim is to consider if P does indeed have mean μ (e.g. is this a fair coin?)

1. Set the significance level (5%/1%/0.1%) used.
2. Perform experiment to find y.
3. Calculate by how many standard deviations y differs from μ, i.e the value $q=(|y-\mu|/\sigma)$
4. Find the probability of a value *at least* q standard deviations away from the mean in $N(\mu,\sigma)$.
5. Reject the claim if this probability is under the significance level being used, i.e. 0.05/0.01/0.001.

Example – Collusion/Cheating in a Game

- Below is very basic single die game.
- There are two players – Alice and Bill – each of whom pay £1 per round.
- A referee – Chris – throws a die twice. The value thrown first is given to Alice. The value thrown second is given to Bill. Only Chris knows *both* values.
- If Alice’s score is higher she wins (£2).
- If Bill’s score is higher then he wins (£2).
- If the scores are equal, Chris gets £1 and the remaining £1 is shared by Alice and Bill.

Example continued

- The die being used is *known* to be fair, i.e. each of the six possibilities are *equally likely* to occur.
- Alice becomes suspicious that the game is “fixed” because of the amount of money she has lost.
- Problem – based on the results how can she be “reasonably confident” she is being cheated as opposed to simply being “unlucky”?

Collecting data for experiment.

- Suppose 120 rounds are played.
- There are 3 possibilities –

1. Alice is *not* being cheated (the game is fair).
2. Chris is cheating *both* players (claiming the same value occurred when it had not).
3. Bill *and* Chris *are colluding* to cheat Alice (Chris wrongly reports that Bill has won or that both players had the same value throw) .

- The data collected by Alice are the amounts that she, Bill, and Chris have after 120 rounds.

Analysing the Outcome

- If the game is being played fairly, then “typically” after 120 rounds we expect –
  A. Alice to have “about” £110.
  B. Bill also to have “about” £110.
  C. Chris to have “about” £20.
- The standard deviation with (A,B) is ≈5.5
- The standard deviation with (C) is ≈4.1

Forming Null Hypotheses

- The Null Hypothesis is that the game is being played fairly. In this case, Alice should have ~£110, (x=110).
- In order to decide *if* she is being cheated, the Normal Distribution with $\mu=110$ and $\sigma=5.5$ – N(110,5.5) is used.

Deciding significance of y (result)

- Statistical studies have shown that
  - $|x-y| \geq 1.65 \times \sigma \Rightarrow$ Significant
  - $|x-y| \geq 2.33 \times \sigma \Rightarrow$ Highly Significant
- $|x-y| \geq 3.09 \times \sigma \Rightarrow$ Very Highly Significant
- The values (1.65,2.33,3.09) apply in *any* experiment in which σ is *known*, for a Normally Distributed population.

Deciding if game was fair – I

- Since σ=5.5 is known, if, after 120 rounds, Alice has won

1. At most £100 i.e. 110-(5.5×1.65),

➢ she is 95% sure she is being cheated.

2. At most £97 i.e. 110-(5.5×2.33),

➢ she is 99% sure she is being cheated.

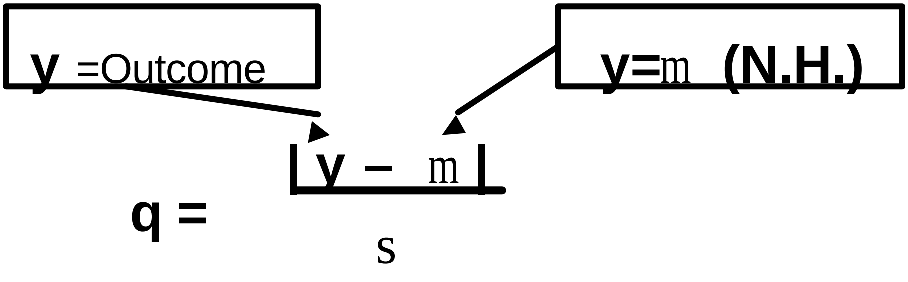
3. At most £93 i.e. 110-(5.5×3.09),

➢ she is 99.9% sure of she is being cheated.

Deciding if game was fair – II

- With μ *claimed* to be 20 and σ *known* to be 4.1, Alice has 95%/99%/99.9% confidence that Chris is cheating if, after 120 rounds, Chris wins
  1. At least £27 i.e. 20+(4.1×1.65)
  2. At least £30 i.e. 20+(4.1×2.33)
  3. At least £33 i.e. 20+(4.1×3.09)

The q-test



y =Outcome

y=m **(N.H.)**

$$\mathbf{q} = \frac{|\,\mathbf{y} - m\,|}{s}$$

q 3 (1.65 ; 2.33 ; 3.09) ⇒ N.H. has *at best* a (20-to-1/100-to-1/1000-to-1) chance of holding

Using the q-test – I

- The main application is in settings studying a collection of outcomes

  Y = { y(1), y(2), … , y(n) }

  sampled from a Normally Distributed population, P, with Standard Deviation σ.
- The Null Hypothesis tested is

  "*The population* P *has mean* x"
- It can be rejected (with some confidence) if y (the *average value* of Y) is "*too many*" standard deviations away from x.

Using the q-test – II

- Assuming that σ has been given (or can be "easily" found), the *only* computations needed are –

1. Collecting the experimental data, (the sample of outcomes, Y).
2. Computing y (the average value of the sample outcomes Y).
3. Comparing y with the "*alleged mean*", x, (presented in the Null Hypothesis).
4. Compute q to determine whether there are grounds to reject the Null Hypothesis.

- Note that *no explicit calculation of probability using* N(x,σ) *is needed*: it's already been done with the 3 significance values – (1.65 ; 2.33 ; 3.09)!

A problem with the q-test

- A major requirement to be satisfied in correct applications of the q-test is that,

  The Standard Deviation (σ) *of the population* (P) *being sampled is known*.
- It is, however, rarely the case with "real" studies, that σ is known or easily found.
- When investigating hypotheses in such situations, the q-test *cannot* be used.

Estimating Standard Deviation I

- Suppose, given a collection

  Y={Y(1), Y(2), … ,Y(n)}

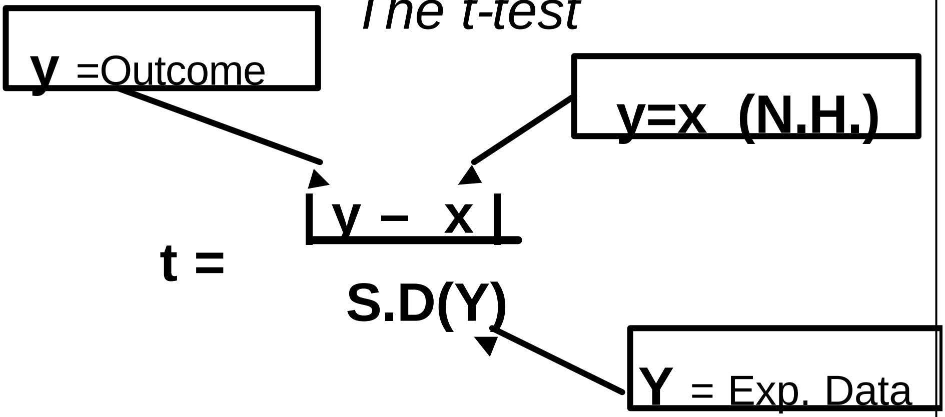
  drawn from a normally distributed population, P, with σ(P) not known, an "*estimate*" of σ(P) can be computed (using Y).
- If this "*estimate*" is "*good enough*" then it could be used in a significance test, other than the q-test.

Estimating Standard Deviation II

- A “natural” starting point is the sample

  Y={Y(1), Y(2), … ,Y(n)}
- This is making the assumption that S.D(Y) ought to provide a “*good enough estimate*” of σ(P).
- We can then replace q (in the q-test) by the number of “estimated deviations” (S.D.(Y)) that y (the experiment outcome) is from x (the *predicted* result).

The t-test



y =Outcome

y=x (N.H.)

$$t = \frac{|y - x|}{S.D(Y)}$$

Y = Exp. Data

t ^{3} (??? ; ??? ; ???) ⇒ N.H. has *at best* a (20-to-1/100-to-1/1000-to-1) chance of holding

Problem – how large must t be?

- The “confidence values” used in the q-test *cannot always be used* in the t-test.
- This is because S.D(Y) is an estimate of σ (the “*true*” standard deviation of P).
- Fortunately, a number of methods are available to compute the value of t that is needed to ensure a given level of significance, e.g. pre-computed tables.
- The number of samples, n, in Y is an important factor in computing this value and is referred to as the “*number of degrees of freedom*”.

Degrees of freedom (d.f.)	5% signif. (number of SD needed).	1% signif. (number of SD needed).	0.1% signif. (number of SD needed).
1	6.31	31.82	318.5
2 … 10	2.92 … 1.81	6.97 … 2.76	22.33 … 4.14
30 … n	1.70 … t(d.f. n)	2.46 … t(d.f. n)	3.39 … t(d.f. n)

More about t–tables I

- The columns give the “number of (estimated) SDs” that a result must differ from the prediction in order for a given level of significance to hold.
- As the “degrees of freedom” value increases, these get smaller, i.e. the larger the sample size (n) the smaller the distance from x has to be to achieve a particular significance level.
- In practice, n=100 (1.66/2.36/3.17) and n=500 (1.65/2.34/3.11) are already “close”, and with n=5000 (1.645/2.327/3.092) are approaching the values used in the q-test.

More about t–tables II

- This last property indicates an important aspect of the so-called “t-distributions”, namely
- Larger population samples (that is, the degrees of freedom) when used to *estimate* standard deviation give *more accurate* estimates.
- As a result t(d.f. n) is “closer to” the confidence limits used in q-tests than t(d.f. n-1).
- If it were possible to use *unlimited* degrees of freedom, then the “estimated” S.D., would in fact be the “*true*” value of σ needed, i.e. the q-test and t-test would be *identical*.

Important applications of the t-test

- There are three main forms –

I. Is a *single* mean (y) “significantly different” from a predicted mean (x)?

II. Given a collection of *paired samples* (from the same population), is there a significant difference between the 2 means?

III. Given *two collections* of samples (from the same population) is there a significant difference between the means? (unpaired)

- We consider (I) and (II) only.

Using the t-test – Single Mean

- We have a collection of outcomes

  Y = { y(1), y(2), … , y(n) }

  sampled from a Normally Distributed population, P.
- The Null Hypothesis tested is

  “*The population* P *has mean* x”
- It can be rejected (with some confidence) if y (the *average value* of Y) is “*too many*” “estimated standard deviations” away from x.

Using the t-test – Paired Samples I

- Suppose we have two methods, Q and R, for solving the same problem, e.g. the two methods for permutations earlier.
- It is often hard to justify claims such as “In general, Q and R are equally good”; or “Q is typically much faster than R”.
- Such claims concern behaviour on average.
- One possibility is to use experiments where the result of comparing Q(Y) with R(Y), when Y is chosen at random.

Example – checking upgrades

- Consider the following scenario:

  The Megahard Corporation releases an upgraded and (allegedly) improved version of its antivirus software. Many clients complain that the upgrade is noticeably slower than its predecessor.

Q. Assuming that all of the complaints about speed come from users with similar hardware (e.g. notebook, laptop) and Operating system, how can Megahard convince its clients that the upgrade is OK?

Example continued

A. Megahard selects a random sample of its clients and invites them to report the results of running both versions of the antivirus software, over a given trial period – the reports are given as pairs of the form

{AV-Old Time ; AV-New Time}

with the time units being, e.g. Millisecond.

Experiment Structure

- Data – the n pairs of run-time data sent.
- Null Hypothesis – the average time taken by AV-New is no worse than that of AV-Old.
- How is a t-test used to analyse this?
- Suppose Megahard recognise that the new system is slower but claim the difference is not significant. They will consider further investigation if there are "*highly significant*" indications that the new version is *slower than* as its predecessor.
- In this case, both 5% and 1% significance levels are relevant: the Null Hypothesis should be *rejected* if the data show a <1-in-100 chance of the (average) run-time of the new version being at most the average run-time of the old version; i.e. a significance level of 1% is used.

Methodology

- Given n pairs {AVO,AVN}

  {<AVO(1),AVN(1)>, …, <AVO(n),AVN(n)>}

define

diff(i) = AVN(i) – AVO(i)

1. Compute M_D, i.e. the average value of the difference between these values (the Null Hypothesis claims this is 0).
2. Compute SD_D an (estimate) of the standard deviation of diff(i).
3. Compute the standard error of the the value M_D using $SE_D = SD_D/n^{0.5}$.
4. Calculated the t-test value as $t=M_D/SE_D$.
5. If t ≥ t(n d.f.) then the Null Hypothesis (that the new version is "no slower") is rejected.

Some features of this experiment

a. Is the *time* taken by an antivirus program its *only* relevant feature? Does it even matter?

b. The second Assessment deals with an alternative experimental comparison of the two antivirus approaches.

Summary

- Statistical significance tests provide methods by which an experimental outcome (y) may be compared with a prediction: the (Null) hypothesis, x.
- The q-test and t-test are 2 examples.
- Important: the outcome of an experiment may *fail to reject* an hypothesis; this *does not imply* the hypothesis is *accepted*.

NOT (Rejected N.H) ^{1} Accepted N.H.

Experiments are not error-proof

- Type I errors

  Rejecting a *true* hypothesis.
- Type II errors

  Not rejecting a *false* hypothesis.
- A "*low*" significance level (5%) can avoid Type II but increase Type I.
- For example, a true hypothesis could fail a 5% test but pass one at 1%
- A "*high*" significance level (0.1%) can avoid Type I but increase Type II.
- For example, a false hypothesis could pass a 0.1% test but fail with one at 1%