Testing Statistical Hypotheses

A: Statistics does not lie - statisticians do!
B: No they don't. They just change significance.

One of the cheif branches of statistical inference is the **testing statistical hypotheses.** Generally, any claim that can be made about population(s) of interest is a *statistical hypothesis*. For example, we hypothesize that:

- The mean μ of the population is 2, or
- Two populations have the same variance, or
- The population is normally distributed, or
- The means of 6 populations are the same, or
- Two populations are independent, etc.

Procedures leading to acceptance[1] or rejection of statistical hypotheses are called statistical tests.

Classical statisticians test hypotheses mainly by using the so-called Neyman-Pearson lemma as a mathematical tool. Because the lemma is very technical and we will not attempt to give it in the introductory course.

In Bayesian framework which will be discussed in Chapter **??**, the testing hypothesis is conceptually straightforward. The hypotheses are assigned probabilities and those with the larger probability are clear winners.

6.1 Notation, errors, p-values, and the power

6.1.1 Choice of H_0

The very first task in testing problems is a formulation of hypotheses. Clearly, there will have at least (and in most of the cases) two competing hypotheses. The hypothesis that we believe is true state of nature is usually denoted by H_0 (null hypothesis) and the competing one is denoted by H_1 (alternative hypothesis).

It is important which of two hypothesis is assigned to be H_0 since the further testing procedure depends on that assignment.

> **Rule:** When our goal is to establish an assertion with substantive support obtained from the sample, the negation of the assertion is taken to be the *null* hypothesis H_0, and the assertion itself is taken to be the alternative hypothesis H_1.

The word *null* in this context can be interpreted to mean that the assertion we seek to establish is actually void.

Examples: Identify the null hypothesis in terms of descriptive statements. The type of answer required is illustrated in the part (a).

(a) The construction engineer wishes to determine if a new cement mix has a better bonding quality than the mix currently in use. The new mix is more expensive, so engineer would not recommend it unless the better quality is supported by the experimental evidence. The bonding quality is to be observed from several cement slabs prepared with the new mix.

---

[1]Statisticians avoid use of the word *accept*. The more careful, but equivalent wording is: *there is not enough statistical evidence to reject*. However in these notes we will use the terms accept and reject, leaving the careful wording to a statistician who may be put in position of being sued for wrong consequences of straightforward wording: accept-reject.

Answer: Null hypothesis H_0: The new mix is not better. Alternative hypothesis H_1: The new mix is better.

(b) A state labor department wishes to determine if the current rate of unemployment in the state varies significantly from the forecast of 6% made two months ago.

Answer: Null hypothesis H_0: The current rate of employment is 6%. Alternative hypothesis H_1: The current rate of employment is different than 6%.

(c) It is claimed that a new treatment is more effective then the standard treatment for prolonging lives of terminal cancer patients. The standard treatment has been in use for a long time, and from records in medical journals the mean survival period is known to be 4.2 years.

Answer: Null hypothesis H_0: The new treatment is as effective as the standerd one. Alternative hypothesis H_1: The new treatment is more effective than the standard one.

(d) Katz et al. (1990) examined the performance of 28 students who answered multiple choice items on the SAT without having read the passages to which the items referred. The mean score (out of 100) was 46.6, with standard deviation 6.8. Random guessing would have been expected to result in 20 correct answers.

Answer: Null hypothesis H_0: The mean score is 20. Alternative hypothesis H_1: The mean score is larger than 20.

(e) A pharmaceutical company claims that its best-selling pain-killer has a mean effective period of at least 4 hours. Do the sample data indicate that the company's claim is to high? Write the null and the alternative hypotheses.

Answer: Null hypothesis H_0: The best-selling pain-killer has a mean effective period of 4 hours. Alternative hypothesis H_1: The The best-selling pain-killer has a mean effective period of less than 4 hours.

6.1.2 Test Statistic, rejection regions and decisions

Suppose we have stated hypotheses H_0 and H_1, and have taken a random sample from the population under research. As in estimation problems, a statistic is calculated from the random sample. Testing is done by looking at the value of the statistic. If the value looks very surprising (in light of the model claimed by H_0, will reject H_0. In other words the data do not support H_0.

We give the following simple example. Pretend we are to test that the first-earned (magic) dime of Scrooge McDuck is fair. Making sure that Miss Magica DeSpell is not arround, Scrooge gives us the dime and we flip it 20 times. We do believe that coin is fair and pose the null hypothesis $H_0 : p = 0.5$, where p is the probability of heads. The expected number of heads in 20 flips is 10, and observed number of heads in neighborhood of 10 (like 8,9,10,11,12) will not surprise us. If we observed 3 heads only, then that will be surprising. The probability of 3 or less heads in 20 flips of a fair coin is only 0.0013. Thirteen times in 10000 flips! So to make a decision we first fix a probability level which is surprising. That is the **level of the test α.**

For some practical reasons the level is chosen from among several typical levels: 0.01, 0.05, 0.10. Chosen level ("surprise level") will be important in deciding to reject or not to reject the null hypothesis.

To introduce a notion of rejection region we have to specify the alternative hypothesis. In the Scrooge dime example we have the three possible alternatives: $H_1 : p \neq 0.5$, $H_1 : p > 0.5$, and $H_1 : p < 0.5$. The first one is called **two-sided** and later two are called **one-sided** hypotheses.

Suppose we fixed $H_1 : p < 0.5$ or "the probability of heads is less than $\frac{1}{2}$." If the test level $\alpha = 0.05$ then observing 5 or less heads will have probability 0.0207 and observing 6 or less heads will have the probability 0.0577 exceeding the level α. Thus observing 5 and less heads is surprising. This defines the **rejection region.** The word *rejection* is connoted to the null hypothesis.

If the observed **test statistic**, in our case the number of heads, is in the rejection region, we reject H_0. Otherwise we fail to reject (or to simply continue to believe in H_0.)

What property of the test was used in making the decision. The test statistics: The number of tails has a Binomial distribution, and the rejection region can be determined to have probability equal or less than the "surprising" level α.

Test Statistic X:number of heads $\quad X \leq 5$: reject H_0
$X > 5$: do not reject H_0.

Plot of rejection region.

6.1.3 Errors in testing

If the null hypothesis is rejected while in fact it is true, then the *error of I kind* is committed. If, on the other hand, we do not reject wrong null hypothesis, then the *error of II kind* is committed (Table 1).

	decide H_0	decide H_1
true H_0	OK	error of I kind
true H_1	error of II kind	OK

It is customary to denote with α the probability of error of the first kind and with β the probability of error of second kind.

A good testing procedure keeps the probabilities of the errors of I and II kind minimal. However, minimizing both errors at the same time, if the sample size is fixed, is impossible. Usually when we minimize α, β increases.

Sometimes in testing problems there is no clear dichotomy: 'established truth' vs 'research hypotheses' and both hypotheses seem to be research hypotheses. For instance, the hypotheses "The new drug is safe" and "The new drug is not safe" are both research hypotheses. In such cases we choose H_0 in such a way that the error of first kind is more serious than the error of second kind. If we choose "The new drug is not safe" as H_0 then the error of first kind is (reject true H_0,i.e. use unsafe drug) is more serious (at least for the patient) than the error of second (accept false H_0, i.e. do not use a safe drug).

That is one reason why α is fixed. We want to control the probability of more serious error. The second convenience of fixing a few values for α are brief statistical tables. Standard values for α are 1%, 5%, and 10%.

6.1.4 p-values

A p-value is the probability of obtaining a value of the test statistics as extreme as or more extreme (in the direction of alternative hypothesis) than that actually obtained, given that the null hypothesis is true.

Adventage of reporting p-values. When researcher reports a p-value as a part of their research findings, readers can set their own level of significance. They can use their own criterium to reject or not reject the null hypothesis, rather than that of the researcher. A lot of information is lost by reporting only that the null hypothesis is rejected at the level $\alpha = 0.05$.

Decisions by p-value:

- If p-value is less than α: reject H_0.
- If p-value is greater than α: do not reject H_0.

6.2 Case study:

6.3 Testing a normal mean when the variance is known: z-test

6.3.1 Power of the z-test

The error of second kind for $H_0 : \mu = \mu_0$, when μ_1 is the actual mean.

- One sided test:

$$\beta = \Phi(z_\alpha - \frac{|\mu_0 - \mu_1|}{\sigma/\sqrt{n}}; \quad \text{Power} = 1 - \beta.$$

- Two sided test:

$$\beta \approx \Phi(z_{\alpha/2} - \frac{|\mu_0 - \mu_1|}{\sigma/\sqrt{n}}; \quad \text{Power} = 1 - \beta.$$

6.4 Testing a normal mean when the variance is unknown: t-test

Theory of t test, unnecessary dichotomy, Splus prog, etc.

The Moon Illusion. Kaufman and Rock (1962) concluded that the commonly observed fact that the moon near the horizon appears larger than does the moon at its zenith (highest point overhead) could be explained on the basis of the greater *apparent* distance of the moon when it is at the horizon. As part of a very complete series of experiments, the authors initially sought to estimate the moon horizon so as to match the size of a standard "moon" that appeared at its zenith, or vice versa. (In these measurements, they used not the actual moon but an artificial one created with special apparatus.) One of the first questions we might ask is whether there really is a moon illusion - that is, whether a larger setting is required to match a horizon moon or a zenith moon. The following data for 10 subjects are taken from Kaufman and Rock's paper and represent the ratio of the diameter of the variable and standard moons. A ratio of 1.00 would indicate no illusion, whereas a ratio other than 1.00 would represent an illusion. (For example, a ratio of 1.50 would mean that the horizon moon appeared to have a diameter 1.50 times the diameter of the zenith moon.) Evidence in support of an illusion would require that we reject $H_0 : \mu = 1.00$ in favor of $H_1 : \mu \neq 1.00$.

Obtained ratio:	1.73	1.06	2.03	1.40	0.95	1.13	1.41	1.73	1.63	1.56

For these data, $N = 10, \bar{X} = 1.463$, and $s = 0.341$.

```
> moon_c(1.73, 1.06, 2.03, 1.40, 0.95, 1.13, 1.41, 1.73, 1.63, 1.56)
> ttest(mu0=1, moon, alt=">")

---------------------------------------------
                  t-test
 Testing H_0: mu= 1  v.s. H_1: mu > 1 .
---------------------------------------------
 :-( Reject H_0.
 p-value= 0.001 is smaller than alpha= 0.05 .
 t-statistic= 4.298 .
 The rejection region cutpoint is (+/-) 1.833 .
```

6.5 Testing the proportion

6.6 Testing the variance

The test for $H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 \neq (<,>)\sigma_0^2$ is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

The rejection region is $\chi^2 < \chi^2_{n-1,\alpha/2}$ and $\chi^2 > \chi^2_{n-1,1-\alpha/2}$ (or for the two one-sided alternatives: $\chi^2 < \chi^2_{n-1,\alpha}$ and $\chi^2 > \chi^2_{n-1,1-\alpha}$.

Aptitude test. Aptitude test should produce scores with a large amount of variation so that an administrator can distinguish between persons with low aptitude and those with large aptitude. The standard test used by a certain university has been producing scores with a standard deviation of 5 points. A new test tried on 20 prospective students produced a sample standard deviation of 8 points. Are the scores from the new test significantly more variable then scores from the standard? Use $\alpha = 0.05$.

6.7 Exercises

Serum Sodium Levels. [2] These data give the results of analysis of 20 samples of serum measured for their sodium content. The average value for the method of analysis used is 140 ppm.

140	143	141	137	132	157	143	149	118	145
138	144	144	139	133	159	141	124	145	139

Is there evidence that the mean level of sodium in this serum is different from 140 ppm?
[$t = \frac{145.55-140}{9.455/\sqrt{20}} = 2.625104$.]

Testing IQ. We wish to test the hypothesis that the mean IQ of the students in a school system is 100. Using $\sigma = 15$, $\alpha = 0.05$ and a sample of 25 students the sample value $\bar{X}$ is computed. For a twosided testing find:

(a) the range of $\bar{X}$ for which we would accept the hypothesis.
(b) If the true mean IQ of the children is 105, find the probability of falsely accepting $H_0 (H_0 : \mu = 100)$.
(c) What are the answers in (a) and (b) if the alternative is onesided, $H_1 : \mu > 100$?

Practice. The following test has been established:

$$H_0 : \mu = 100 \quad vs \quad H_1 : \mu > 100$$

The population standard deviation is $\sigma = 15$. The probabilities of Type I and Type II errors are

$$P(\text{Type I error}) = 0.01, \quad P(\text{Type II error}) = P(H_0|\mu = 110) = 0.15$$

.

Find the critical value and sample size. State the decision rule.

[2] *Source:* National Quality Control Scheme, Queen Elizabeth Hospital, Birmingham, referenced in *Data* by D. F. Andrews and A. M. Herzberg, Springer, 1985.

Just Practice. A random sample of size $n = 50$ from a population produced the mean $\bar{X} = 3.1$ and sample standard deviation $s = 0.5$. Suppose that your objective is to show that population mean μ exceeds 3.

(a) State the hypotheses H_0 and H_1.

(b) Perform a test at level $\alpha = 0.05$.

(c) Explain why can you use normal distribution tools even when you know that X is non-normal?

Testing Piaget. Two sets of elementary school students are taught mathematics by two different methods: classical (Group 1) and small group interactive teaching by discovery based on Piagetian theory (Group 2). The results of a learning test are analyzed to test the difference in mean scores using the two methods. Group 1 has 10 students with a mean score of 64.1 and a standard deviation of 5.1; group 2 has 8 students with a mean score of 68.0 and a standard deviation of 5.9.

Test the hypothesis that methods have no influence on test scores against the alternative that the method applied to the group 2 is better. Take $\alpha = .05$.

Bricks. A purchaser of bricks suspects that the quality of the bricks is deteriorating. From past experience, the mean crushing strength of such bricks is 400 pounds, with a standard deviation of 20 pounds. A sample of 100 bricks yielded a mean of 395 pounds. The purchaser believes that standard deviation is the same. Test the hypothesis that the mean quality has not changed against the alternative that it was deteriorated. Choose $\alpha = .05$. (*Hint:* Assume one sided alternative.)

```
> ttest(mu0=400, x=395, sig=20, n=100, alt="<")

---------------------------------------------
                 t-test
 Testing H_0: mu= 400  v.s. H_1: mu < 400 .
---------------------------------------------
 :-( Reject H_0.
 p-value= 0.007 is smaller than alpha= 0.05 .
 t-statistic= -2.5 .
 The rejection region cutpoint is (+/-) 1.66 .
```

Soybeans. According to advertisements, a strain of soybeans planted on soil prepared with a specified fertilizer treatment has a mean yield of 500 bushels per acre. Fifty farmers, who belong to a cooperative, plant the soybeans. Each uses a 40-acre plot and records the mean yield per acre. The mean and variance for the sample of 50 farms are $\bar{x} = 485$ and $s^2 = 10,045$.

Use p-value for this test to determine whether the data provide sufficient evidence to indicate that the mean yield for the soybeans is different from the advertised.

```
> ttest(mu0=500, x=485, sig=sqrt(10045), n=50)

---------------------------------------------
                 t-test
 Testing H_0: mu= 500  v.s. H_1: mu != 500 .
---------------------------------------------
 :-) Do not reject H_0.
 p-value= 0.295 is larger than alpha= 0.05 .
 t-statistic= -1.058 .
 The rejection region cutpoint is (+/-) 1.677 .
```

Great white shark. One of the most feared predators in the ocean is the great white shark *Carcharodon carcharias*. Although it is known that the white shark grows to a mean length of 21 feet (record: 39 feet), a marine biologist believes that the great white sharks off the Bermuda coast grow much longer due to unusual feeding habits. To test this claim, a number of full-grown great white sharks are captured off the Bermuda coast, measured and then set free. However, because the capture of sharks is difficult, costly, and very dangerous, only three are sampled. Their lengths are 24, 20, 22 feet.

1. Do the data provide sufficient evidence to support marine biologist's claim? Use $\alpha = 0.1$.
2. What assumptions must be made in order to carry out the test?

Public Health. A manager of public health services in an area downwind of a nuclear test site wants to test the hypothesis that the mean amount of radiation in the form of Strontium-90 in the bone marrow (measured in picocuries) for citizens who live downwind of the site does exceed that of citizens who live upwind from the site. It is known that "upwinders" have a mean level of Strontium-90 of 1 picocurie. Measurements of Strontium-90 radiation for a sample of $n = 16$ citizens who live downwind of the site was taken, giving $\bar{X} = 3$. The population standard deviation is $\sigma = 4$.

a) Test the (research, alternative) hypothesis that downwinders have a higher Strontium-90 level than upwinders. Assume normality and use a significance level of $\alpha = 0.05$.
(i) State H_0 and H_1
(ii) State the appropriate test statistic
(iii) Determine the critical region of the test
(iv) State your decision
(v) What would constitute a type-II error in this setup? *Describe in less than 20 words.*

Drill. Suppose $\bar{X}$ is the mean of a normal random sample of size 25, where $\sigma = 10$. Test $H_0 : \mu = 75$ versus $H_1 : \mu < 75$ by rejecting the null hypothesis when $\bar{X} \leq 71.08$.

(a) Find α.
(b) Find β for $\mu = 73$.

Drill. 5) Test $H_0 : \mu = 15$ versus $H_1 : \mu \neq 15$ if a normal random sample of size 20 gives $\bar{X} = 9.6$ and $s = 4.7$. Let $\alpha = 0.05$.

Till now we have the fixed sample size n. Often we are in a situation to choose n (design the experiment). For example, it may be up to us to decide how many respondents to interview in a pool. We designed sample sizes in interval estimation to achieve given precision and confidence level.

As an example of designing a sample size in a testing setup, consider a problem of testing $H_0 : \mu = m_0$ using $\bar{X}$ of a sample size n. Let the alternative has fixed value μ_1, i.e. $H_1 : \mu = \mu_1.(> \mu_0)$ How large n should be so that the power $1 - \beta$ is 0.90 and $\alpha = 0.05$?

The critical region is $\bar{X} > \mu_0 + 1.645 \cdot \frac{\sigma}{\sqrt{n}}$. We want the power $= P(\bar{X} > \mu_0 + 1.645 \cdot \frac{\sigma}{\sqrt{n}} | \mu = \mu_1) = 0.90$, i.e.

$$P(\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} > \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + 1.645) = 0.9.$$

Since $P(Z > -1.282) = 0.9$, it follows $\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} = 1.282 - 1.645 => n = \frac{8.567 \cdot \sigma^2}{(\mu_1 - \mu_0)^2}$

More generally, if we want to achieve the power $1 - \beta$ within the significance α in testing $H_0 : \mu = \mu_1$, we need $n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{\mu_1 - \mu_0}$ observations.

Remarks:

(1). If σ^2 is unknown, the above is just an approximation for sample size. However, the approximation is very good if n is large. Exact calculations will include quantiles of t distribution

(2). For two sided alternatives α is replaced by $\alpha/2$.

> The sample size for fixed α, β,
>
> $$n = \frac{\sigma^2}{\Delta^2}(z_\alpha + z_\beta)^2.$$
>
> If the alternative is two sided z_α is replaced by $z_{\alpha/2}$. In that case the sample size is approximate.
> If σ^2 is not known, substitution of an estimate will give an approximate sample size.

Binomial Model. In testing the null hypothesis $H_0 : p = .6$ vs. the alternative $H_1 : p < .6$ for a binomial model, the rejection region of a test has the structure $X \leq c$, where X is the number of successes in n trials. For each of the following tests, determine the level of significance and the probability of a type II error at the alternative $p = .3$.
(a) $n = 10, c = 2$
(b) $n = 10, c = 3$
(c) $n = 20, c = 7$

Pens. A company claims their pens will write for over 100 hours. If we take this statement to apply to the mean μ, show how to state H_0 and H_1 in a test designed to establish the claim.

More Drills. (a) What is the form of the rejection region when testing $H_0 : \mu \leq 102$ against $H_1 : \mu > 102$ based on a sample size $n = 50$?

(b) If $\bar{x} = 103$ and $s = 6$, what is the conclusion of your test with $\alpha = .05$? Also find the significance probability and interpret the result.

STAT Cola again. A sample of 70 cans of cola was analyzed for amount of caffeine. The results yielded $\bar{x} = 57.8$ milligrams and $s = 7.3$. Test

$$H_0 : \mu \geq 60 \quad vs. \quad H_1 : \mu < 60$$

with $\alpha = .10$.

Cancer therapy. Researchers in cancer therapy often report only the number of patients who survive for a specified period of time after treatment rather than the patients' actual survival times. Suppose that 30% of the patients who undergo the standard treatment are known to survive 5 years. A new treatment is administered to 100 patients, and 38 of them are still alive after a period of 5 years.

(a) Formulate the hypotheses for testing the validity of the claim that the new treatment is more effective than the standard therapy.

(b) Test with $\alpha = .05$ and state your conclusion.

Dwarf Plants. A genetic model suggests that 80% of the plants grown from a cross between two given strains of seeds will be of the dwarf variety. After breeding 200 of these plants, 136 were of the dwarf variety.

(a) Does this observation strongly contradict the genetic model?

(b) Construct a 95% confidence interval for the true proportion of dwarf plants obtained from the given cross.

Sally on a highway. Sally is stranded on a highway awaiting the arrival of an AAA truck to repair her car. The local AAA club says that it responds to emergency calls in an average of 35 minutes with a standard deviation of 4.2 minutes. After speaking to 30 friends who were in similar situations, Sally finds that the AAA responded to calls in an average of 39 minutes. Does this indicate that the average response time is more than 35 minutes? (Use a 5% level of significance.)

Fast balls. A baseball manager claims that the average speed of a fast ball thrown by a particular pitcher is 93 MPH with a standard deviation of 10.71 MPH. Fifty fast balls thrown by the pitcher are randomly selected. If the average speed of these pitches is found to be 91 MPH, should we reject the baseball manager's claim? (Use a 10% level of significance.)

Registration Process. The administration of Farliey College claims that the entire registration process should last an average of 40 minutes with a standard deviation of 6 minutes. A reporter for the college newspaper claims that the average time is much more than 40 minutes. In a sample of 49 students, it is found that the average time needed by these students to complete the registration process was 47 minutes. Can we accept the administration's claim? (Use a 5% level of significance).

Computer Games. Certain computer games are thought to improve spatial skills. A Mental Rotations Test, measuring spatial skills, was administered to a sample of school children after they had played one of two types of computer game. Construct 95% confidence intervals based on the following mean scores, assuming that the children were selected randomly and that the Mental Rotations Test scores have a normal distribution in the population.

(a) After playing the "Factory" computer game: $\bar{X} = 22.47, s_x = 9.44, n = 19$.

(b) After playing the "Stellar" computer game: $\bar{X} = 22.68, s_x = 8.37, n = 19$.

(c) After playing no computer game (control group): $\bar{X} = 18.63, s_x = 11.13, n = 19$.

Burnout of nurses. Researchers (Keane, Ducette, and Adler, 1985) investigated the extent of "burnout" among nurses in various hospital settings. Levels of such feelings as "I often think about finding a new job" or "I frequently get angry with my patients" were ascertained by questionnaire and measured on a Staff Burnout Scale for Helth Professionals. The study focused on whether burnout levels for Intensive Care Unit (ICU) nurses differed from those for other hospital nurses. For all nurses in the large hospital studied, the burnout scale had a mean of 52.1. Among a sample of 25 ICU nurses the mean was 49.9, with a standard deviation of 14.3. Assuming that the burnout scale has an approximately normal distribution, test the hypothesis that the mean burnout level of ICU nurses is different from that for this hospital's nurse population as a whole. Include these steps.

(a) Formally state the null and alternative hypotheses, and explain whether a one- or two-sided test is appropriate.

(b) Calculate the degrees of freedom. What critical value of t is required to reject H_0 at $\alpha = .05$?

(c) Calculate the t statistic.

(d) Summarize your conclusions in one paragraph.

Eggs in nest. The average number of eggs laid per nest per season for the Eastern Phoebe bird is quantity of interest. A random sample of 70 nests was examined and the following results were obtained.

Number of Eggs/Nest	1	2	3	4	5	6
Frequency f	3	2	2	14	46	3

Test the hypothesis that the true average number of eggs laid per nest by the Eastern Phoebe bird is equal to 5 vs twosided alternative. Use $\alpha = 0.05$.

Penguins. Penguins are popular birds, and the king penguin *Aptenodytes patagonica* is the most popular penguin of all. Researcher is interested in testing that a mean height of king penguins from a small island is less than known mean height $\mu = 36$ in. for the whole penguin population. The random measurements of height of 12 adult birds are

34	34	37	33	36	35	33	35	36	33	35	37

State the hypotheses and perform the test at the level $\alpha = 0.05$.

Sol. $\bar{X} = 34.83, s = 1.467, t = -2.76$, *p-value=0.00923,* $t_{0.95,11} = 1.795$.

Aniline. Organic chemists often purify organic compounds by a method known as *fractional crystallization.* An experimenter wanted to prepare and purify 4.85 grams of aniline. Ten 4.85 quantities of aniline were individually prepared and purified to acetanilide.

(a) Test the hypothesis that the mean dry yield is less than 4 grams if the sample mean yield observed was $\bar{X} = 3.85$. Population variance $\sigma^2 = 0.08$ is assumed known and $\alpha = 0.05$.

(b) Report the p-value.

(c) For what values of $\bar{X}$ the null hypothesis will be rejected (α is still 0.05)?

(d) Define and explain the power of a test (No more than 4 sentences). What is the power of the test if the alternative is $H_1 : \mu = 3.6$.

$(a) z = -1.677051$, p-value =0.0475.

Jigsaw. An experiment with a sample of 18 nursery-school children involved the elapsed time required to put together a small jigsaw puzzle. The times were:

3.1	3.2	3.4	3.6	3.7	4.2	4.3	4.5	4.7
5.2	5.6	6.0	6.1	6.6	7.3	8.2	10.8	13.6

(The sample mean is 5.783, and the sample standard deviation is 2.784).

(a) Calculate a 95% confidence interval for the population mean.

(b) Test the hypothesis $H_0 : \mu = 5$ against the two sided alternative. Take $\alpha = 10\%$.

Close Encounters. Otis[3] (1979) interviewed moviegoers waiting to see the *space aliens* film "Close Encounters of the Third Kind." Each moviegoer was asked to state his or her degree of agreement with the statement "Life on Earth is being observed by intelligent aliens," on a scale from 1 (strongly disagree) to 5 (strongly agree). Assume that the population standard deviation is known: $\sigma = 1$.

1. Why $H_0 : \mu = 3$ is a natural null hypothesis?

2. The purpose of the study was to test Otis' assertion that individuals selected movies that they were predisposed to believe. Formulate Otis' assertion as an H_1 hypothesis.

3. The test adopted can be described as follows:

If the sample mean of $n = 25$ moviegoers is larger than 3.4, reject H_0.

(a) Find α for the test.

(b) Find β against $H_1 : \mu = 4$.

(c) If the observed $\bar{X}$ for $n = 25$ was in fact 3.5, what is p-value?

Anxiety A psychologist has developed a questionnaire for assessing levels of anxiety. The scores on the questionnaire range from 0 to 100. People who obtain scores of 75 and more are classified as *anxious*. The questionnaire has been given to a large sample of people who have been diagnosed with an anxiety disorder, and scores are well described by a normal model with a mean of 80 and a standard deviation of 5. When given to a large sample of people who do not suffer from an anxiety disorder, scores on the questionnaire can also be modelled as normal with a mean of 60 and a standard deviation of 10.

What is the probability that the psychologist missclassifies a non-anxious person as anxious?

What is the probability that the psychologist erroneously labels a truly anxious person as non-anxious?

Rats and Mazes. Eighty rats selected at random were taught to run a new maze. All of them finally succeeded in learning the maze, and the number of trials to perfect the performance was normally distributed with a mean of 15.4. Long experience with population of rats trained to run a similar maze shows that the number of trials to success is normally distributed with a mean of 15 and a standard deviation of $\sigma = 2$. Is the new maze harder for rats to learn than the older one? Assume $\alpha = 0.01$.

---

[3] Otis, L. (1979). Selective exposure to the film "Close Encounters". *Journal of Psychology, 101, 293-295.*

6.8 Splus Programs

6.8.1 z test and its use

```
> ztest <-
  function(mu0 = 0, x, sig, n = length(x), alt = "!=", alpha = 0.05, r = 3)
{
        cat("\n--------------------------------------------\n")
        cat("                 z-test                    \n")
        cat(" Testing H_0: mu=", mu0, " v.s. H_1: mu", alt, mu0, ".\n")
        cat("--------------------------------------------\n")
        zi <- (mean(x) - mu0)/(sig/sqrt(n))
        p <- 2 * pnorm( - abs(zi))
        if(alt == "<")
                p <- pnorm(zi)
        if(alt == ">")
                p <- 1 - pnorm(zi)
        if(alpha > p)
               cat(" :-( Reject H_0.\n p-value=", round(p, r),
                      "is smaller than alpha=", alpha, ".\n z-statistic=",
                      round(zi, r),
                     ".\n The rejection region cutpoint is (+/-)", round( -
                      qnorm(alpha), r), ".\n")
        else cat(" :-) Do not reject H_0.\n p-value=", round(p, r),
                     "is larger than alpha=", alpha, ".\n z-statistic=",
                      round(zi, r),
                     ".\n The rejection region cutpoint is (+/-)", round( -
                     qnorm(alpha), r), ".\n")
}
```

6.8.2 t-test and its use

```
> ttest <- function(mu0 = 0, x, sig = sqrt(var(x)), n = length(x),
 alt = "!=", alpha = 0.05, r = 3)
{
        cat("\n----------------------------------------------\n")
        cat("                  t-test                     \n")
        cat(" Testing H_0: mu=", mu0, " v.s. H_1: mu", alt, mu0, ".\n")
        cat("----------------------------------------------\n")
        ti <- (mean(x) - mu0)/(sig/sqrt(n))
        p <- 2 * pt( - abs(ti), n - 1)
        if(alt == "<")
                p <- pt(ti, n - 1)
        if(alt == ">")
                p <- 1 - pt(ti, n - 1)
        if(alpha > p)
              cat(" :-( Reject H_0.\n p-value=", round(p, r),
                    "is smaller than alpha=", alpha, ".\n t-statistic=",
                    round(ti, r),
                    ".\n The rejection region cutpoint is (+/-)", round( -
                    qt(alpha, n - 1), r), ".\n")
        else cat(" :-) Do not reject H_0.\n p-value=", round(p, r),
                    "is larger than alpha=", alpha, ".\n t-statistic=",
                    round(ti, r),
                    ".\n The rejection region cutpoint is (+/-)", round( -
                    qt(alpha, n - 1), r), ".\n")
}
```