Institute for the Advancement of University Learning
&
Department of Statistics

Descriptive Statistics for Research
(Hilary Term, 2002)

Lecture 7: Hypothesis Testing

(I.) Introduction

An important area of statistical inference involves using observed data to decide between competing explanations, or hypotheses, about the process under investigation. Statisticians refer to this aspect of inference as "hypothesis testing." The following example illustrates the reasoning that underlies an hypothesis test.

> **Example***:* According to Hacking (1965:75), the first person to publish an account of the reasoning behind a statistical inference was the Scottish physician John Arbuthnot in his paper "an Argument for Divine Providence taken from the constant Regularity of the Births of Both Sexes," communicated to the Royal Society in 1710. The hypothesis of interest is that there is an even chance of a child being born male or female. Arbuthnot studied the register of births for the city of London between 1629 and 1710 and found that for every single year in that period the number of males christened exceeded the number of females. He argued that if there were an even chance for male and female births, the distribution of births should be like outcomes from tossing a fair coin.
>
> Let H denote the hypothesis that the probability of a male birth in any particular year is equal to ½, and let M denote the event of a year having more male births than female. If H were true, then the observed data would have a probability of $(½)^{82}$ which, to quote Arbuthnot, "will be found easily by the table of logarithms to be 1/4836000000000000000000000." He went on to say that if an event happened in London, as reported by the registers, and if H were true, then the chance of the observed event (i.e. every year in the period producing more male christenings) would be minute. Arbuthnot did not note that his argument is valid only if we are solely interested in the *number* of *M*s that occur and do not care about the *order* in which these years occurred. If we do consider the order in which the *M*s occurred, *any* result (e.g., having 41 *M*s, each of them followed by a year with a majority of female births) has exactly the same probability (i.e., $(½)^{82}$). Arbuthnot concluded: "it follows that it is art, not Chance, that governs" in the

> distribution of sexes, hence, the title of his paper. As Fisher (1973:42) says in a similar context: "either an exceptionally rare chance has occurred, or the theory of random distribution is not true."

This type of reasoning forms the cornerstone of statistical hypothesis testing.

As we have stated many times before, we can test hypotheses regarding general parameters (e.g., the mean), parameters that are model quantities (e.g., the rate, λ, of a Poisson distribution), or the structure of a population (e.g., "is the population normally distributed?"). In addition, we can use either parametric or nonparametric procedures to construct such tests, and the comments made in previous lectures regarding which approach to take apply here as well. For example, parametric tests require us to make assumptions about the structure of the underlying population distribution, whereas nonparametric tests do not; parametric tests perform 'better' if the assumed population distribution is correct; and, nonparametric tests are applicable in a broader range of situation, but are 'conservative.'

(II.) The Logic of Hypothesis Testing

According to Rice (1995: 299), statistical hypothesis testing is a formal means of distinguishing between probability distributions on the basis of random variables generated from one of the distributions. The general idea can be written as follows:

- *Prior* to observing data:
    1. State a baseline hypothesis: This hypothesis is usually a statement of 'no change' or of maintaining the status quo. Hence, we often refer to this conjecture as the ***null hypothesis***;
    2. State an ***alternative hypothesis***: This is typically the hypothesis of interest to the researcher. That is, it is the hypothesis that the researcher wishes to demonstrate to be true;

- *After* observing data:
    1. Decide how likely the observed data is, assuming the null hypothesis to be true;
    2. Reject the null hypothesis in favour of the alternative if there is 'enough evidence' to suggest doing so. Otherwise, do not reject the null hypothesis.

This procedure can be likened to a criminal trial under an adversarial legal system, like the one in place in the UK. Under this system, a person accused of a crime is assumed innocent, and the onus is placed on the prosecution to 'prove' guilt. We can think of the assumption of innocence as the null hypothesis, and the suspicion of guilt as the alternative (of course, this means the experimenter is the prosecution – who, or what, do you think is the defence?). Data is presented to the court in the form of evidence, and it is the court's job to weigh the evidence thus presented. If the evidence is sufficiently strong, the accused is found guilty. If not, the assumption of innocence is upheld.

However, this analogy raises an interesting question. If the evidence presented is insufficient to convict, does this mean that the accused is truly innocent? Although we would like to think so, unfortunately, due to the way the process is constructed, the answer is that we do not know. All we can really say is that there was insufficient evidence to convict.

This is also the case with hypothesis testing. When the data provide insufficient evidence in favour of the alternative hypothesis, we are unable to say that the null hypothesis is 'true.' Rather, in statistical terms, we say that we "cannot reject the null in favour of the alternative" or that we "fail to reject the null." It is incorrect to say that "the null hypothesis is true."

To illustrate the hypothesis testing decision process, consider the following problem.

> **Example:** Suppose we are given a sample $X_1, X_2, \cdots, X_n$ of n = 15 heights from a population of males. Further, suppose that we assume the population distribution of these heights to be normal with known variance σ^2=81 cm^2. In addition, suppose we know this population has a mean height of either μ_0=175 cm or μ_1=180 cm. Figure 1 shows this situation.

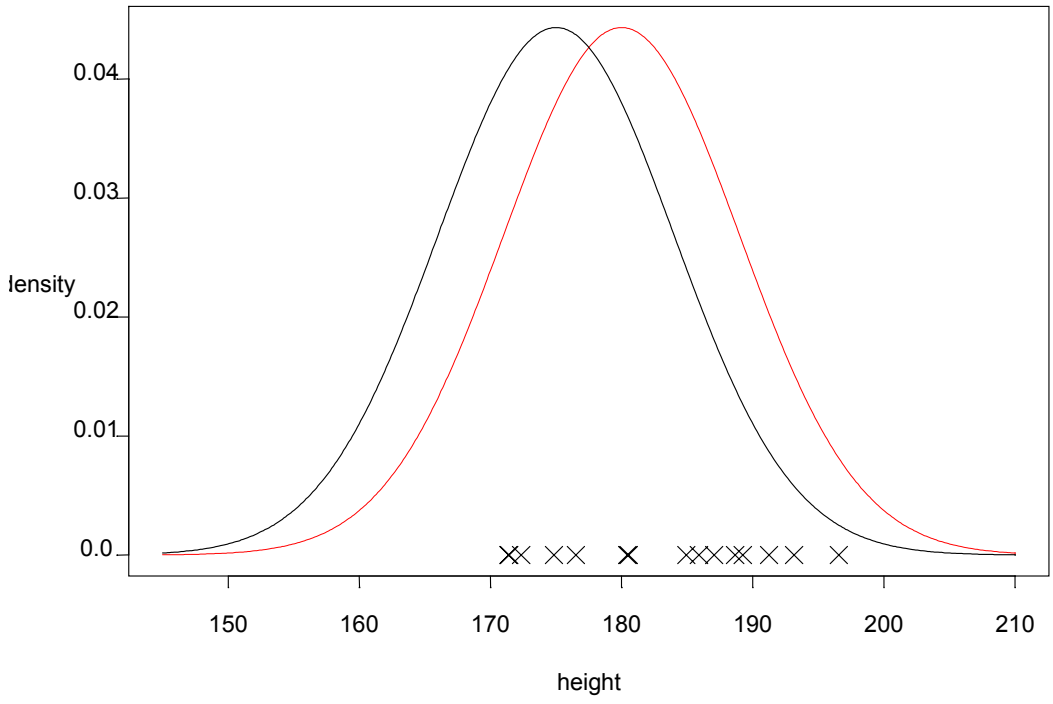

Figure 1: *Two simple hypotheses with data observations (n = 15)*

Note that

a) Both hypotheses about the mean values refer to models for the *population*, and they are usually expressions of competing theories;

b) The hypotheses are stated *before* observing the sample values;

c) A decision is reached *after* analysing the sample values.

How do we decide which of these two possible distributions ($N(175,81)$ or $N(180,81)$) explains the data better? The most common approach has already been described above and is known formally as the Neyman-Pearson paradigm[1] of inference. In the following sections, we examine each part of the process in turn, beginning with the construction of hypotheses.

(III.) The Null and Alternative Hypotheses

As we have seen, the first step in our inferential procedure is to divide the probability distributions into the ***null hypothesis*** (denoted by H_0) and the ***alternative hypothesis*** (H_1, though some texts use H_A). The usual notation for a hypothesis test is H_0 vs H_1. In the present situation we could write

$$H_0 : \mu = 175 \quad \text{vs.} \quad H_1 : \mu = 180 .$$

In Arbuthnot's example, we could say that H_0 specifies a binomial distribution with parameters (n=82, π=0.5) and that H_1 corresponds to a binomial distribution with parameters (n=82, π>0.5). For the moment, we will regard these two classes of hypotheses as mere labels. Later, however, we will explore a relationship between them that arises as a consequence of the Neyman-Pearson paradigm.

As has already been alluded to, from a logical standpoint the only decisions available to us in a statistical test of hypotheses are the rejection of the null hypothesis or a failure to reject it. Again, note that "failure to reject" is not synonymous with "acceptance"! To quote Fisher (1973:45), "A test of significance contains no criterion for 'accepting' a hypothesis. According to circumstances it may or may not influence its acceptability." In this lecture, however, we will use the colloquial expressions 'accepting a hypothesis' and 'rejecting a hypothesis' simply to indicate whether the data provides evidence for or against the hypothesis, respectively.

[1] After Jerzy Neyman and Egon Pearson, who first published an account of this theory of inference in 1933.

Simple and Composite Hypotheses

If a hypothesis completely specifies the distribution to which it refers (i.e., if it specifies the pdf family to which the distribution belongs as well as the values of all parameters required for that family), it is called ***simple***. Both hypotheses in the previous example are simple; this would not be the case if their common variance were unknown. If the distribution related to a hypothesis is not completely specified, the hypothesis is called ***composite***. An example of a composite hypothesis would be the use of the alternative hypothesis $H_1 : \mu \neq 175$ in the heights example.

One– and Two – sided Alternatives

Often, hypotheses specify the value of a parameter or a range of values for a parameter. When an alternative hypothesis is composite (that is, when it specifies a parameter range as opposed to a specific value), it can be one of two types: ***one – sided*** or ***two – sided***. For instance, in the heights example we might be interested either in the alternative hypothesis H_1: μ > 175 or in the alternative hypothesis H_1: μ ≠ 175; these alternatives are examples of one-sided and two-sided alternative hypotheses, respectively. As we will see, whether a hypothesis is one– or two- sided plays an important role in the testing of hypotheses.

(IV.) Testing the Hypotheses

According to the Neyman-Pearson paradigm, a ***test*** is a rule for deciding whether or not to reject H_0 in favour of H_1. This decision is based on the evidence contained in our sample, $\mathbf{X} = \{X_1, X_2, \cdots X_n\}$. How we go about extracting such evidence from the sample depends on the problem at hand. In general, though, we construct a ***test statistic***, denoted $T(\mathbf{X})$, from the data and under the assumption that H_0 is true; note that $T(\mathbf{X})$ is a function of the data and that, as such, it is a random variable and has a sampling distribution. Once $T(\mathbf{X})$ has been constructed, it is then used in the decision making process: sufficiently 'unlikely' values of $T(\mathbf{X})$ lead us to reject our assumption of H_0 in favour of the alternative. To illustrate, consider again the heights example. Here, a suitable test statistic would be the sample mean. If we observe a value of $\overline{X}$ suitably close to 175 cm, we would conclude that the data support H_0; the bigger the difference between the observed value of $\overline{X}$ and 175, the stronger the evidence against H_0.

Acceptance and Rejection Regions

How do we decide what values of $T(\mathbf{X})$ are 'likely' and what values aren't? In order to answer this question, we first define two important quantities.

- ***Acceptance region*** of a test*:* The set of values of $T(\mathbf{X})$ for which we would be unprepared to reject H_0;

- ***Rejection region*** (often also called the ***critical region***) of the test: The set of values of $T(\mathbf{X})$ for which we would be prepared to reject H_0 in favour of H_1.

Obviously, these two quantities, but especially the rejection region, play a crucial role in hypothesis testing. We will see how to find the rejection region later, but we should note here that this region is constructed using the sampling distribution of the test statistic, in much the same way as we saw with confidence intervals. In some situations (including the ones with which we will deal in this lecture), we can use statistical theory to find this distribution analytically. When this is not possible, we must use other techniques, such as simulation. The general idea is illustrated graphically in Figure 2.

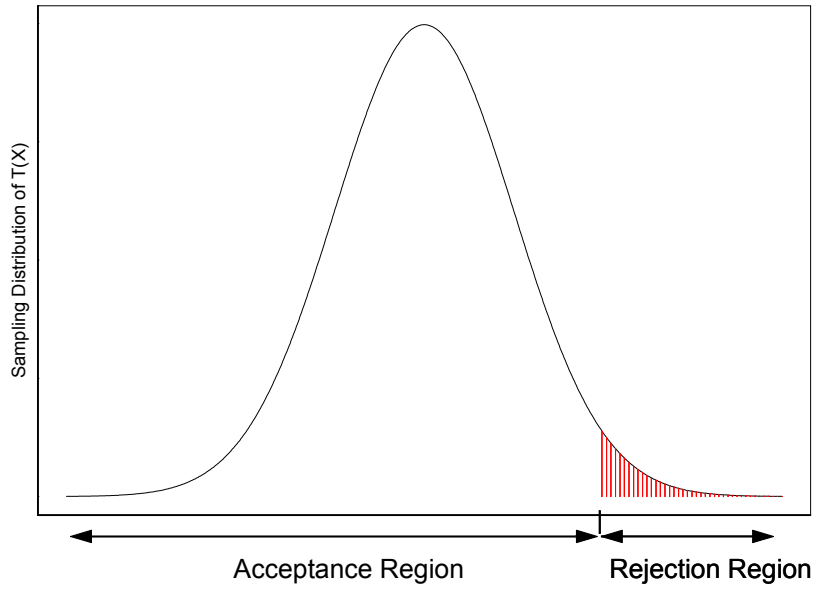

***Figure 2**: An example critical region for a hypothesis test based on T(X)*

(V.) Types of Error

A re-examination of Figure 1 shows that it would be unlikely for the data to come from the model specified by the null hypothesis (i.e., a normal distribution with mean 175 cm). However, it is still possible that the null hypothesis is actually true and that, due only to bad luck, we observed a sample of 15 individuals who are mostly from the upper part of the null distribution; thus, we do not want to be too hasty in rejecting the null hypothesis. This point illustrates how when we consider testing, for example, $H_0 : \mu = 175$ vs. $H_1 : \mu > 175$, we might incur two types of error:

- ***Type I error***: *H_0 may be rejected when it is true.* We denote the probability of committing this type of error by α. If H_0 is simple, then α is called the ***significance level*** of the test. If H_0 is composite, then the probability of a type I error depends

on which particular member of H_0 is true; in that case, the significance level is defined as the maximum (actually supremum) of these probabilities. In our legal analogy, a type I error can be thought of as a finding of guilt when the defendant was, in fact, innocent.

- ***Type II error***: H_0 *may be accepted when it is false.* The probability of committing this type of error is denoted **β**. If H_1 is composite, β depends on which particular member of H_1 holds. Again, from our legal analogy, we may think of this type of error as the event that the defendant is found innocent when he/she is actually guilty.

The probability that H_0 is rejected when it is false is called the ***power*** of the test, and is equal to 1 - β. The power of a test measures how sensitive the test is at detecting deviations from H_0. The following table illustrates both types of errors.

		Population (hypotheses)	
		H_0	H_1
Sample (decision)	accept H_0	Correct	type II error
	reject H_0	type I error	correct

Table 1*: Types of error in hypothesis tests*

We would like to construct tests with α and β as small as possible. Indeed, since they are probabilities of error, we would like them, ideally, to be equal to 0. However, this can only be achieved in trivial examples of no practical value. In practice, given a fixed sample size, there exists a trade-off between these two quantities: in order to decrease α, we must increase β and vice versa. The Neyman-Pearson paradigm solves this conflict by imposing an *asymmetry* between H_0 and H_1: the significance level, α, is fixed in advance, usually at a rather small number (e.g., 0.05), and then an attempt is made to construct a test with the smallest possible value for β.

We will illustrate these concepts with the heights example.

Example (continued): Let X denote the height of an individual. Suppose we have a sample of size n = 15, and consider the following simple hypotheses:

$$H_0 : X \sim N(175, 81) \text{ vs. } H_1 : X \sim N(180, 81).$$

This can be written, equivalently, as

$$H_0 : \mu = 175 \text{ vs. } H_1 : \mu = 180,$$

provided we make suitable assumptions (normality and known population variance equal to 81).

The sample mean, $\overline{X}$, is a plausible test statistic for this problem since we know, for example, that it is an unbiased estimator of μ and that, under suitable conditions, it has a normally distributed sampling distribution.

Recall that the critical region is defined as those values of the test statistic, here $\overline{X}$, for which we would not reject H_0. It is obvious that, given the form of our hypotheses, the larger the observed value of $\overline{X}$, the more evidence there is against H_0. Therefore, a rule that specifies this test says that we should reject H_0 if $\overline{X} > k$, where k is a number arbitrarily chosen as a cut off. Note the following points:

- The larger k is, the more difficult it becomes to reject H_0, and thus the more the probability of incorrectly rejecting H_0, assuming it is true (i.e. α), decreases.

- The smaller k is, the easier it is to reject H_0, and the probability of not rejecting it, assuming that H_1 is true (i.e. β), decreases.

Figure 3 illustrates the changes in α and β as a function of k. The value of α is calculated under the assumption that H_0 is true, and is simply the probability of $\overline{X} > k$ under this assumption. The distribution of the test statistic obtained by assuming H_0 to be true is called the ***null distribution***. In our example, the null distribution is $N(175, 81/15)$, the distribution of $\overline{X}$ assuming the true mean is 175 cm. In Figure 3, the distributions shown are the null distribution (continuous line) and the distribution under the alternative (dotted line). The value of α is the area shaded with horizontal lines; the value of β is the area shaded with vertical lines.

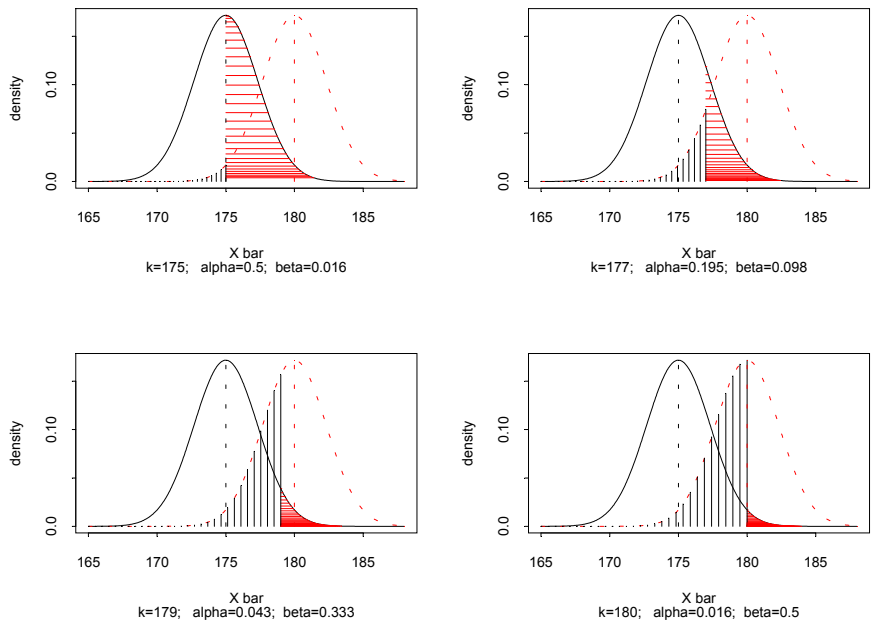

Figure 3: *Changes in* α *and* β *for* $H_0 : X \sim N(175, 81)$ vs $H_1 : X \sim N(180, 81)$ with *n=15*

This figure illustrates the trade-off between α and β: we see that increasing *k* decreases α, but increases β. If we use a boundary too close to the value specified by H_0, it will be easy to decide incorrectly against the null hypothesis; for instance, for k=175, α has a high value regardless of the sample size. On the other hand, if we choose a boundary too close to the value defining H_1, it will be difficult to reject the null hypothesis (either correctly or incorrectly) so α tends to 0 and β increases to ½. This trade-off is also evident in the following table, which shows changes in α and β for different sample sizes and values of *k*.

	n=30		N=60		n=150	
k	**α**	**β**	**α**	**β**	**α**	**β**
175	0.500	0.001	0.500	0.000	0.500	0.000
177	0.112	0.034	0.043	0.005	0.003	0.000
179	0.007	0.271	0.000	0.195	0.000	0.087
180	0.001	0.500	0.000	0.500	0.000	0.500

Table 2: *Trade - off between* α *and* β *for increasing k, for sample sizes of 30, 60, and 150*

Lastly, looking at the above table, we should note that, for a given *k*, β decreases as the sample size increases; thus, from a purely statistical point of view, it is desirable to have a larger sample.

The Neyman – Pearson Lemma

Up until this point, we have fixed the value of *k* and calculated α (and β). However, as described above, the Neyman-Pearson paradigm tells us that we need to fix α and

calculate k (and β). We can fix α to be arbitrarily small enough, but what can we do to make sure that the corresponding β is the smallest possible, given α and n? For the heights example, we could consider other plausible test statistics (e.g. the median or the trimmed mean), or perhaps some transformation of $\overline{X}$. The Neyman-Pearson lemma[2] gives a criterion[3] which allows us to construct 'optimal' tests in the sense that they minimise β given α and n. We won't go into the mathematical details: a lucid account of the theory appears, for example, in Rice (1995:§9.3). Most of the tests that will be used in this course are based on the Neyman-Pearson lemma, and thus we will not have to worry about controlling β.

(VI.) The General Approach to Hypothesis Testing

A General Recipe

Based on the ideas discussed above, we are now in a position to describe a general 'recipe' for constructing statistical hypothesis tests. The procedure can be written as follows:

- *Before* data collection/observation:

  a. State the hypotheses H_0 and H_1;
  b. Choose and fix the significance level of the test (α);
  c. Establish the critical region of the test corresponding to α. This region depends both on the null distribution of the test statistic T and on whether the alternative hypothesis is one- or two- sided.

- *After* data collection/observation:

  d. Calculate the value of T from the data sample - call this value t^*;
  e. Compare t^* with the null distribution of T in order to see whether or not it falls in the critical (rejection) region;
  f. Make a decision about the hypotheses.

We demonstrate the approach with the heights example.

> **Example (continued):** It can be shown that the test statistic for an optimal test of $H_0 : X \sim N(\mu_0, \sigma^2)$ vs $H_1 : X \sim N(\mu_1, \sigma^2)$, with σ^2 known, is
>
> $$T = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}},$$
>
> where $\overline{X}$ is the sample mean. Recall that if $X \sim N(\mu, \sigma^2)$, then

[2] A *lemma* is a mathematical proposition which has been demonstrated to be true

[3] Called the L*ikelihood Ratio*

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1).$$

In the present case, we know that $\overline{X} \sim N\left(\mu, \sigma^2/n\right)$ since the variable X is assumed to have a *normal* distribution. Therefore, the null distribution of T (i.e., the sampling distribution of the test statistic assuming H_0 to be true) is *N(0,1),* assuming σ is known.

Consider testing the hypotheses $H_0 : \mu = 175$ vs. $H_1 : \mu > 175$. Our first step (after specifying the hypotheses) is to choose and fix the significance level: here we will take $\alpha = 0.05$.

The next step is to determine the critical region of the test. We have a simple null hypothesis against a composite alternative, and this alternative hypothesis is one-sided. Thus, the critical region will be of the form $T > k$, since the bigger T is, the more evidence against the null hypothesis. That is, we choose k such that

$$P\left(T > k \mid H_0 \text{ true}\right) = 0.05,$$

or, alternatively,

$$P\left(T \leq k \mid H_0 \text{ true}\right) = 0.95.$$

Since we have stated above that the null distribution is *N(0,1)* in this example, we know from the previous lecture that k corresponds to the 0.95th quantile of the null distribution, which is $z_{1-\alpha} = z_{0.95} = 1.645$ from standard normal tables. This boundary for the critical region guarantees, by the Neyman-Pearson lemma, the smallest value of β obtainable for the given values of α and n.

Suppose that now we obtain a sample of $n = 15$ heights from this population, and that the observed sample mean for this data is 182.9 cm. The value of the test statistic is thus

$$t^* = \frac{182.9 - 175}{\sqrt{81/15}} \approx 3.39.$$

Figure 4 shows the right-hand side of a *N(0,1)* distribution with the critical region for T (defined by $T > 1.645$ for $\alpha = 0.05$) as well as the observed value of the test statistic.

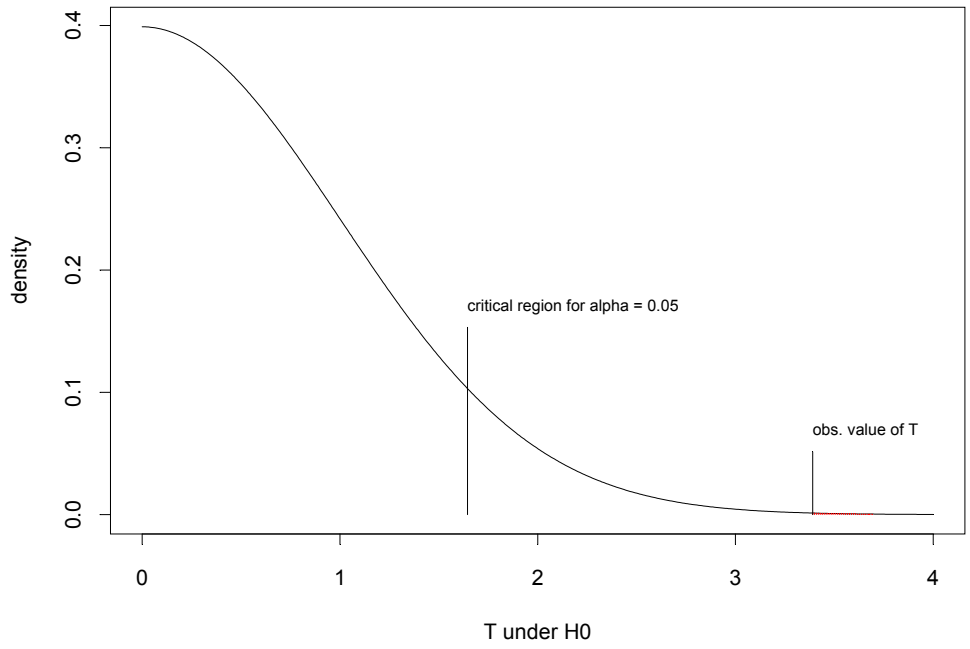

***Figure 4**: Critical region for α = 0.05 and observed value of T*

We can see that the observed value of T falls in the critical region for the test. **We therefore reject H_0 in favour of H_1 at the α = 0.05 level of significance, and conclude that $\mu > 175$.**

Note that the alternative hypothesis refers only to the parameter μ being larger than 175 cm; it does not specify a particular value for μ. We could have performed a test of two simple hypotheses if we were interested in a particular value for the alternative, *before* data collection.

Alternative Hypotheses and Critical Regions

We mentioned previously that the type of (composite) alternative hypothesis (i.e., whether it is one-sided or two-sided) affects the calculation of the critical region of a test. In this section we elaborate on this idea.

In the previous example, note that after choosing and fixing α, we calculated the *lower bound* of the rejection region as that point *above which* $100\times\alpha$ % of probability lay, under the null distribution. We did this because we were interested in an alternative hypothesis regarding values of μ greater than 175 cm (i.e., a one-sided alternative hypothesis).

This is an example of an ***upper- tailed test***. Obviously (perhaps under different circumstances), we could also be interested in testing a hypothesis about values of μ less than 175cms. This is an example of a ***lower – tailed test***. Finally, when we don't have an

opinion either way, but are simply interested in whether μ is different from μ_0, we have what is called a ***two –tailed test***. Upper – and lower – tailed tests correspond to one – sided alternative hypotheses; two – tailed tests correspond to two – sided alternative hypotheses.

The critical regions we calculate depend on what type of test we are carrying out. For an upper – tailed test, we calculate *k* such that

$$P(T > k) = \alpha .$$

That is, $k_{upper} = T_{1-\alpha}$, the $(1-\alpha)$th quantile of the null distribution.

For a lower – tailed test, we calculate *k* such that

$$P(T < k) = \alpha .$$

That is, $k_{lower} = T_{\alpha}$, the α th quantile of the null distribution. Note that, for the same level of significance, $k_{lower} = -k_{upper}$ for symmetric null distributions.

For a two – tailed test, we calculate *k* such that

$$P(|T| > k) = \alpha .$$

That is, the rejection region is split into two sub – regions: one defined by values of T less than –*k* , and one defined by values of T greater than *k*, where $k = T_{1-\alpha/2}$, the $\left(1-\frac{\alpha}{2}\right)$th quantile of the null distribution. These ideas are illustrated graphically in Figure 5.

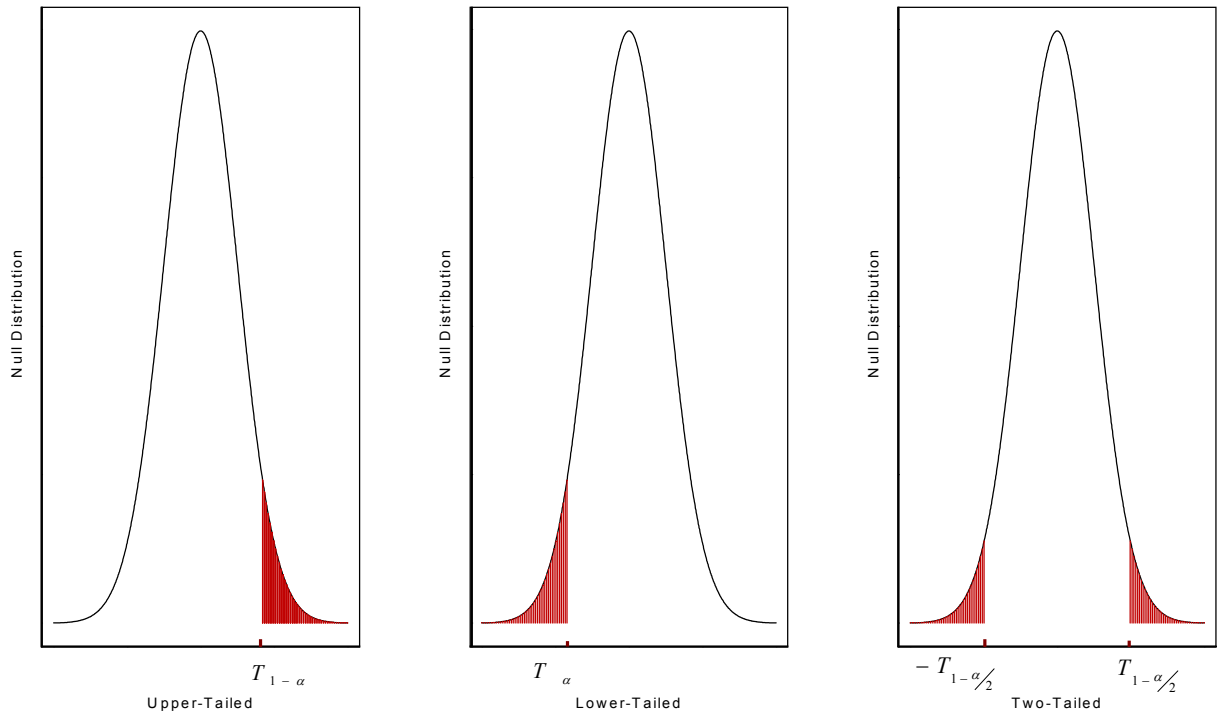

***Figure 5**: Critical regions for various types of test*

(VII.) P - values

The final concept associated with tests of hypotheses is the ***p*-value.** The p - value is defined as the smallest value of α for which the null hypothesis would be rejected, given the data.

We can see from Figure 4 that, for certain values of α less than 0.05, we would have still rejected H_0. This holds true for values of $T < t^* = 3.39$. If we now apply the definition of the *p*-value, we see that

$$p - value = P\left[T \geq t^*\right].$$

That is, the p – value of a test is the probability of observing, by chance, a value of the test statistic as extreme as, or even more extreme than, the one we did observe, assuming that the null hypothesis were true. If this probability is extremely small, then *either* H_0 holds and we have observed an extremely rare event, *or* H_0 is 'false' (there is sufficient evidence to reject H_0 in favour of the alternative). Thus, the *p*-value can be seen as a measure of the 'risk' taken when, assuming H_0 is true, we decide to reject this hypothesis. If this 'risk' is 'sufficiently small,' we can feel confident that we are not observing a freak random event; rather, we are seeing strong evidence against the null hypothesis. We define 'sufficiently small' to be values of p less than the level of significance of the test, α.

Example (continued): Recall that the observed value of our test statistic was $t^* = 3.39$. We find the *p*-value, using statistical tables (or, more commonly, computer software) to be 0.00035, or 3.5 in 10,000. This is much smaller than 0.05, and we conclude that it is safe to reject H_0 in favour of the alternative, given the evidence contained in the data. The *p*-value is the area under the null distribution curve to the right of t^* in Figure 4.

As with rejection regions, how we calculate *p* - values depends on the type of test we have constructed. For an upper – tailed test, we have already seen the definition of the p – value:

$$p - value = P\left[T \geq t^*\right].$$

For a lower – tailed test, we have

$$p - value = P\left[T \leq t^*\right],$$

and for a two – tailed test we have

$$p - value = P\left[\,|T| \geq t^*\right].$$

This implies that a two – sided *p* – value is twice as large as its one – sided counterpart.

In many applications, it makes more sense to report a *p* - value than to merely report whether or not H_0 was rejected for a particular α. By reporting a *p* – value, we get an idea of how strong the evidence of the data is: smaller *p* – values imply stronger evidence against the null hypothesis. Most computer packages return the (two – sided) *p* - value as the final result of a hypothesis test, often under the (dubious and confusing) title 'significance level.'

(VIII.) The Relationship Between Hypothesis Tests and Confidence Intervals

There is a direct correspondence between hypothesis tests and confidence intervals. Simply put, a 100(1-α)% confidence interval for a parameter contains all the values of that parameter for which the null hypothesis of a test would not be rejected at the α level of significance. Therefore, a hypothesis test can be performed as described above, or we can construct a confidence interval for the parameter of interest and see if the hypothesised value (under the null) falls in this interval. If it does, then we do not have sufficient evidence to reject the null. For instance, in the heights example, we could use the techniques of Lecture 5 to construct a 95% confidence interval for μ; if this confidence interval did not contain the value 175 cm, then we would reject H_0 against the two-sided alternative $H_1 : \mu \neq 175$ at the α=0.05 level. Before proceeding, we should note that we could instead use the appropriate one-sided 95% confidence interval (of the form

[*lowerbound*,∞]) to test H_0 against the one- sided alternative $H_1 : \mu > 175$ at the α=0.05 level.

(IX.) Incorrect Uses of Hypothesis Tests: Multiple Tests and Data Driven Hypotheses

Multiple Tests

Suppose we are testing a hypothesis regarding many parameters. For example, suppose we had data on the crop yield of four different types of fertilizer. We might want to test the hypothesis that the mean yield for all four fertilizer types is equal, versus the alternative that at least one of the treatment means differs from the others. (We will see how to do this in Lecture 8.)

Suppose that we carried out such a test at the 0.05 level of significance, and rejected the null in favour of the alternative. Now what do we do? We would obviously like to know which treatment gives the highest crop yield. In order to do this, we could compare all pairs of fertilizers to find which fertilizer gives significantly higher yields. There are 6 pairwise comparisons, and thus 6 tests to perform.

Recall that α is the probability of rejecting the null when it is in fact true. When we set this probability to, for example, 0.05, what we are saying is that if we were to repeat the test 20 times, we would expect that on one of these occasions we would make a type I error. Thus, when we use the same data to perform multiple comparisons (tests), we need to be aware that we are increasing the chance of drawing spurious conclusions. That is, we are increasing the chance of concluding that H_0 is true when it is, in fact, not.

We can avoid this problem by making each individual comparison more conservative. That is, for each individual test, we make its significance level (α_{ind}) smaller than α in order to maintain an ***overall significance level*** of α. Note that α, the overall significance level, refers to the probability that *at least one* of the multiple null hypotheses tested will be rejected when they are all actually true, whereas α_{ind} refers to the probability that any individual null hypothesis will be rejected when it is true. There are many different methods that can be used to decide on a value for α_{ind}, but perhaps the most straightforward is the technique known as the ***Bonferroni correction.*** Simply, if we wish to make *m* tests on the same data at an overall significance level of α, we should set the significance level of each test at

$$\alpha_{ind} = \frac{\alpha}{m}.$$

Data Driven Hypotheses

Another commonly occurring mistake when undertaking hypothesis tests is to generate hypotheses based on the results of other hypothesis tests on the same data. Such situations arise as follows. Suppose that before data collection, our alternative hypothesis was two -

sided. After data collection, we carried out the test and found that the *p*-value was, for example, 0.08. This is larger than 0.05, so we cannot reject the null hypothesis at the 0.05 level of significance. However, we know that if we had chosen a one - sided alternative hypothesis, the *p*-value would have been half that observed under the two – sided alternative, and thus given us a significant result. We therefore construct another (one – sided) test, and obtain a significant result.

Unfortunately, we should not do this for several reasons. First, it is equivalent to peeking in a 'pin the tail on the donkey' competition. Second, it violates the principles of the hypothesis test. Notice that we very carefully stated that hypotheses were to be specified *before* looking at the data (ideally, even before collecting the data). Finally, we run into problems with multiple comparisons, as described above.

The basic point to remember is that we specify hypotheses *before* we observe the data. If a particular test reveals something interesting about the process under investigation, and hence generates another hypothesis, ideally we should conduct another experiment to test this subsequent hypothesis.

(X.) Other Types of Hypothesis Tests

Recall that, above, we stated that hypothesis tests can be used to compare competing theories about either the parameters of a population or the structure of a population. Additionally, we noted that hypothesis tests can be either parametric or nonparametric. In this lecture, we have just seen an example of a parametric test of two hypotheses about a population parameter (μ). In this section, however, we will discuss nonparametric tests and tests of structural properties.

(X.a.) Parametric and Nonparametric Tests

A parametric test requires that precise assumptions about the population distribution of the quantity of interest be satisfied in order to use it. Nonparametric tests do not. The above example of testing the population mean of heights is an illustration of a parametric test since the test relies on the assumption that $\overline{X} \sim N(\mu, \sigma^2/n)$. In our heights example, this assumption about the distribution of $\overline{X}$ was satisfied because the underlying random variable, *X*, was assumed to be normally distributed. Note, however, that this assumption about the distribution of $\overline{X}$ would also be satisfied even if *X* were not assumed to be normal as long as the sample size, *n*, were sufficiently large (by the CLT).

The parametric vs. nonparametric nomenclature can be confusing since both classes of tests refer to parameters. The reason for this labelling is a technical one: usually, nonparametric tests are applied to parameters such as the median, which, although they are parameters in the broad sense of the term, do not, in general, define a distribution. On the other hand, parameters such as the mean or the variance often characterise a distribution. For example, the normal distribution is uniquely specified by its mean and variance.

Whenever possible, parametric tests are to be preferred to nonparametric ones because, as long as the assumptions required by the parametric tests are (more or less) satisfied, parametric tests have a larger power, (1-β), than their nonparametric analogues. However, situations can arise in which the specific assumptions required by a parametric test do not hold, thereby invalidating the use of the parametric test; in these cases, a nonparametric test is often preferable. A good introduction to nonparametric tests is the text by Sprent (1993).

(X.b.) Goodness of Fit Tests

Instead of testing hypotheses regarding a population parameter, we may be interested in comparing two hypotheses about the *structure* of the underling population. For instance, in the heights example above, we were told to assume that the data were sampled from a normally distributed population. However, we might be interested in formally testing this assumption using the data in our sample. In this section, we discuss two procedures for testing hypotheses about the distribution of a random variable. These tests are both nonparametric since they do not assume a specific distributional form for the data (obviously, since the distribution is what is being tested!).

(X.b.i.) Chi-square Goodness of Fit Test

This test is appropriate for testing hypotheses regarding the distribution of *discrete* random variables only. The idea behind this test is very simple: once the range of the random variable has been divided into classes of values (perhaps consisting of just one value per class if the range contains only a small number of values), the observed frequencies of the classes (in the data sample) are compared to what we would expect to see under the assumption that the probability model specified by the null hypothesis is true. The test statistic is a measure of the distance between the observed and expected frequencies. The value of this statistic is then compared with the critical region for the test (which is constructed from the null distribution).

Suppose that we observe a sample of size n and that we obtain the frequencies for m different values or classes. The test statistic is then

$$\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

where O_i and E_i are the observed and expected frequencies for the i^{th} class, respectively. The 'closer' the observed frequencies are to the expected frequencies for each class, the less likely we are to reject H_0. If there are discrepancies between the data and the expected frequencies under H_0, then χ^2 will be large and we are more likely to reject the null.

The null distribution of χ^2 is called the ***chi-square distribution*** (hence the name of the test). Its only parameter, *df*, is called the ***degrees of freedom***. For χ^2, the number of degrees of freedom is

$$df = m - 1 - p.e.$$

where *p.e.* is the number of parameters estimated in order to calculate the expected frequencies E_i. The critical region of the test consists of 'large' values of χ^2, as they indicate discrepancies between the data and the hypothesised model. For this test, the p-value is the probability that a chi-square distribution with *df* degrees of freedom will take a value larger than or equal to the calculated value of χ^2; this probability can be easily calculated using a chi-square table in any basic statistics test or using a statistics software package.

Example: Consider the data studied by Richardson (1944) on the temporal distribution of wars. Richardson examined the period 1500 to 1931, and counted the number of outbreaks of war per year. The r.v. of interest, *X*, counts the number of wars occurring in each year of that 432 year period. The observed frequencies indicate, for instance, that there were 223 years with 0 outbreaks of war.

The null hypothesis is that this data is a realization of a Poisson distributed random variable (i.e., the random variable, *X*, has a Poisson distribution). We decided on this distribution for the null hypothesis since it is a good model for rare random events occurring at a constant rate.

In order to calculate the expected frequencies under the null hypothesis, we need to estimate the parameter λ of the Poisson distribution for this data. Using the sample mean to estimate λ, we obtain an estimate of 0.6921. The expected frequencies are then found by plugging this value for λ into the probability mass function formula for the Poisson distribution (Lecture 2) for each category (X), and multiplying these probabilities by the total number of events (432). For example, the expected frequency when X = 0 is given by

$$E_1 = 432 \times \frac{0.6921^0 \times e^{-0.6921}}{0!}.$$

The observed and expected frequencies are:

X	**0**	**1**	**2**	**3**	**4**	**≥ 5**
observed	223	142	48	15	4	0
expected	216.23	149.65	51.79	11.95	2.07	0.28

The agreement between the data and the values produced by the model is very good. The value of χ^2 is 0.1047, on 6 – 1 – 1 = 4 degrees of freedom (since we had to

estimate one parameter - the mean of the Poisson distribution), and the associated *p*-value is 0.99. That is, we would be almost certainly wrong to reject the null hypothesis of a Poisson model for these data. One implication of not rejecting the hypothesis is that war emerges randomly and with a small constant rate over a large period of time. This, in turn, implies that there are a large number of potential conflicts, each having a small probability of becoming a war in any given year.

(X.b.ii.) Kolmogorov-Smirnov Goodness of Fit Test

This test also measures the discrepancy between the data and a hypothesised distribution. To see how, we first have to define the ***empirical distribution function (edf)***. This function, evaluated at any value x, is the proportion of observations in the sample which are less than or equal to x.

The step function shown in Figure 7 is the *edf* for the heights from a sample of 199 British men. The dotted line is the (probability) cumulative distribution function of a normal random variable with parameters (i.e., with mean and variance) equal to the appropriate corresponding sample estimates (i.e., the sample mean and sample variance for the 199 heights). As we can see from the proximity of the two functions in Figure 7, the normal distribution appears to be a good model for these data.

The Kolmogorov-Smirnov statistic, *KS*, is defined as the maximum of the differences between the edf and the cumulative distribution function specified by the null hypothesis. The rejection region for *KS* clearly corresponds to large values of the statistic. We should note that the distribution of this statistic is the same regardless of which distribution is hypothesised for the random variable.

For the example shown in Figure 7, the observed value of *KS* is 0.05, which yields a *p*-value of 0.5. Therefore, we cannot reject the null hypothesis that the data comes from a normally distributed population.

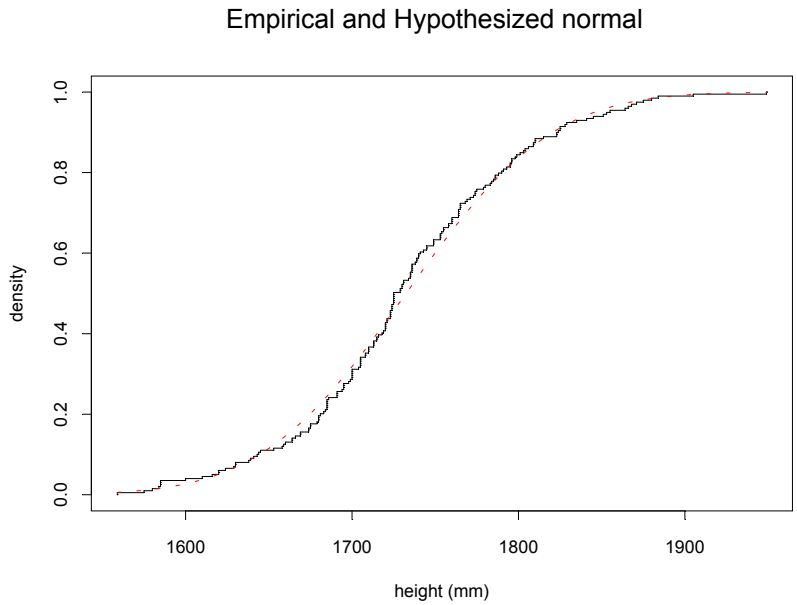

***Figure 7**: Comparison of edf and normal cdf for 199 heights*

(XI.) References

Fisher, RA (1973). *Statistical Methods and Scientific Inference, (3rd edition).* Hafner Press, New York.

Hacking, I (1965). *Logic of Statistical Inference*, Cambridge University Press, Cambridge.

Hand, DJ and CC Taylor (1987). *Multivariate Analysis of Variance and Repeated Measures*: *a practical approach for behavioural scientists.* Chapman & Hall, London.

Henkel, RE (1976). *Tests of Significance*. Sage University Papers series in Quantitative Applications in the Social Sciences, series no. 4. Sage Publications, Beverly Hills and London.

Howell, DC (1992). *Statistical Methods in Psychology (3rd edition).* PWS-Kent, Boston.

Hsu, JC (1996). *Multiple Comparisons: Theory and Practice.* Chapman and Hall, London.

Rice, JA (1995). *Mathematical Statistics and Data Analysis.* Duxbury Press, Belmont, California.

Richardson, LF (1944). "The distribution of wars in time", *Journal of the Royal Statistical Society,* **107**, 242-250.

Sprent, P (1993). *Applied Nonparametric Statistical Methods (2nd edition).* Chapman & Hall, London.

MCB (I-2000), *KNJ* (III-2001), *JMcB* (III-2001)