Notes on Sampling and Hypothesis Testing

Allin Cottrell*

1 Population and sample

In statistics, a *population* is an entire set of objects or units of observation of one sort or another, while a *sample* is a subset (usually a proper subset) of a population, selected for particular study (usually because it is impractical to study the whole population). The numerical characteristics of a population are called *parameters*. Generally the values of the parameters of interest remain unknown to the researcher; we calculate the "corresponding" numerical characteristics of the sample (known as *statistics*) and use these to *estimate*, or make inferences about, the unknown parameter values.

A standard notation is often used to keep straight the distinction between population and sample. The table below sets out some commonly used symbols.

	size	mean	variance	proportion
Population:	N	μ	σ^2	π
Sample:	n	$\bar{x}$	s^2	p

Note that it's common to use a Greek letter to denote a parameter, and the corresponding Roman letter to denote the associated statistic.

2 Properties of estimators: sample mean

Consider for example the sample mean,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

If we want to use this statistic to make inferences regarding the population mean, μ, we need to know something about the probability distribution of $\bar{x}$. The distribution of a sample statistic is known as a *sampling distribution*. Two of its characteristics are of particular interest, the mean or expected value and the variance or standard deviation.

What can we say about $E(\bar{x})$ or $\mu_{\bar{x}}$, the mean of the sampling distribution of $\bar{x}$? First, let's be sure we understand what it means. It is the *expected value* of $\bar{x}$. The thought experiment is as follows: we sample repeatedly from the given population, each time recording the sample mean, and take the average of those sample means. It's unlikely that any given sample will yield a value of $\bar{x}$ that precisely equals μ, the mean of the population from which we're drawing. Due to (random) *sampling error* some samples will give a sample mean that exceeds the population mean, and some will give an $\bar{x}$ that falls short of μ. But if our sampling procedure is *unbiased*, then deviations of $\bar{x}$ from μ in the upward and downward directions should be equally likely. On average, they should cancel out. In that case

$$E(\bar{x}) = \mu = E(X) \tag{1}$$

or: the sample mean is an *unbiased estimator* of the population mean.

So far so good. But we'd also like to know how *widely dispersed* the sample mean values are likely to be, around their expected value. This is known as the issue of the *efficiency* of an estimator. It is a comparative

[*] Last revised 2002/01/29.

concept: one estimator is more efficient than another if its values are more tightly clustered around its expected value. Consider this alternative estimator for the population mean: instead of $\bar{x}$, just take the average of the largest and smallest values in the sample. This too should be an unbiased estimator of μ, but it is likely to be more widely spread out, or in other words less efficient than $\bar{x}$ (unless of course the sample size is 2, in which case they amount to the same thing).

The degree of dispersion of an estimator is generally measured by the standard deviation of its probability distribution (sampling distribution). This goes under the name *standard error*.

2.1 Standard error of $\bar{x}$

What might the standard error of $\bar{x}$ look like? In other words, what factors are going to influence the degree of dispersion of the sample mean around the population mean? Without giving a formal derivation, it's possible to understand intuitively the formula:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{2}$$

The left-hand term is read as "sigma sub x-bar". The sigma tells us we're dealing with a standard deviation, and the subscript $\bar{x}$ indicates this is the standard deviation of the distribution of $\bar{x}$, or in other words the standard error of $\bar{x}$. On the right-hand side in the numerator we find the standard deviation, σ, of the population from which the samples are drawn. The more widely dispersed are the population values around their mean, the greater the scope for sampling error (i.e. drawing by chance an unrepresentative sample whose mean differs substantially from μ). In the denominator is the square root of the sample size, n. It makes sense that if our samples are larger, this reduces the probability of getting unrepresentative results, and hence narrows the dispersion of $\bar{x}$. The fact that it is $\sqrt{n}$ rather than n that enters the formula indicates that an increase in sample size is subject to diminishing returns, in terms of increasing the precision of the estimator. For example, increasing the sample size by a factor of four will reduce the standard error of $\bar{x}$, but only by a factor of two.

3 Other statistics

We have illustrated so far with the sample mean as an example estimator, but you shouldn't get the idea that it's the only one. For example, suppose we're interested in the *proportion* of some population that has a certain characteristic (e.g. an intention to vote for the Democratic candidate). The population proportion is often written as π. The corresponding sample statistic is the proportion of the sample having the characteristic in question, p. The sample proportion is an unbiased estimator of the population proportion

$$E(p) = \pi \tag{3}$$

and its standard error is given by

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} \tag{4}$$

Or we might be particularly interested in the variance, σ^2, of a certain population. Since the population variance is given by

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

it would seem that the obvious estimator is the statistic

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

But actually it turns out this estimator is biased. The bias is corrected in the formula for *sample variance*:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{5}$$

(with a bias-correction factor of $\frac{n}{n-1}$).

4 The shape of sampling distributions

Besides knowing the expected value and the standard error of a given statistic, in order to work with that statistic for the purpose of statistical inference we need to know its *shape*. In the case of the sample mean, the Central Limit Theorem entitles us to the assumption that the sampling distribution is Gaussian—even if the population from which the samples are drawn does not follow a Gaussian distribution—provided we are dealing with a large enough sample. For a statistician, "large enough" generally means 30 or greater (as a rough rule of thumb) although the approximation to a Gaussian sampling distribution may be quite good even with smaller samples.

Here's a rather striking illustration of the point. Consider, once again, the distribution of X = the number appearing uppermost when a fair die is rolled. We know that this distribution is not close to Gaussian: it's rectangular. But recall what the distribution looked like for the average of the two face values when two dice are rolled: it was triangular. What happens if we crank up the number of dice further? The triangle turns into a bell shape, and if we compute the distribution of the mean face value when rolling five dice it already looks quite close to the Gaussian (see Figure 1).

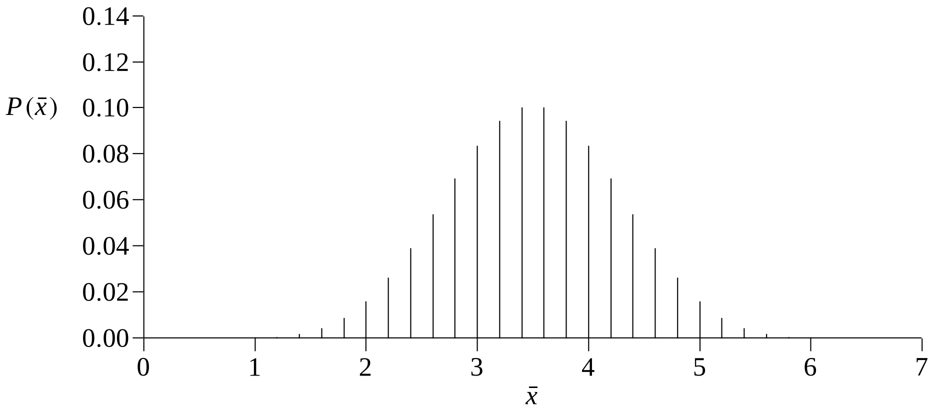
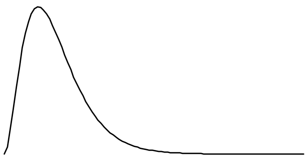


Figure 1: Distribution of mean face-value, 5 dice

We can think of the graph in Figure 1 as representing the sampling distribution of $\bar{x}$ for samples with $n = 5$ from a population with $\mu = 3.5$ and a rectangular distribution. Although the "parent" distribution is rectangular the sampling distribution is a fair approximation to the Gaussian.

Not all sampling distributions are Gaussian. We mentioned earlier the use of the sample variance as an estimator of the population variance. In this case the ratio $(n-1)s^2/\sigma^2$ follows a skewed distribution known as χ^2, with $n-1$ degrees of freedom (below).



Nonetheless, if the sample size is large the χ^2 distribution converges towards the normal.

5 Probability statements, confidence intervals

If we know the mean, standard error and shape of the distribution of a given sample statistic, we can then make definite probability statements about the statistic. For example, suppose we know that $\mu = 100$ and $\sigma = 12$ for a certain population, and we draw a sample with $n = 36$ from that population. The standard error of $\bar{x}$ is $\sigma/\sqrt{n} = 12/6 = 2$, and a sample size of 36 is large enough to justify the assumption of a Gaussian sampling distribution. We know that the range $\mu \pm 2\sigma$ encloses the central 95 percent of a normal distribution, so we can state

$$P(96 < \bar{x} < 104) \approx .95$$

That is, there's a 95 percent probability that the sample mean lies within 4 units (= 2 standard errors) of the population mean, 100.

That's all very well, you may say, but if we already knew the population mean and standard deviation, then why were we bothering to draw a sample? Well, let's try relaxing the assumptions regarding our knowledge of the population and see if we can still get something useful. First, suppose we don't know the value of μ. We can still say

$$P(\mu - 4 < \bar{x} < \mu + 4) \approx .95$$

That is, with probability .95 the sample mean will be drawn from within 4 units of the unknown population mean. So suppose we go ahead and draw the sample, and calculate a sample mean of 97. If there's a probability of .95 that our $\bar{x}$ came from within 4 units of μ, we can turn that around: we're entitled to be 95 percent confident that μ lies between 93 and 101. That is, we can draw up a 95 percent *confidence interval* for μ as $\bar{x} \pm 2\sigma_{\bar{x}}$.

There's a further problem though. If we don't know the value of μ then presumably we don't know σ either. So how can we compute the standard error of $\bar{x}$? We can't, but we can *estimate* it. Our best estimate of the population standard deviation will be s, the standard deviation calculated from our sample. The *estimated standard error* of $\bar{x}$ is then

$$s_{\bar{x}} \equiv \hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{6}$$

(The "hat" or caret over a parameter indicates an estimated value.)

We can now reformulate our 95 percent confidence interval for μ: $\bar{x} \pm 2s_{\bar{x}}$. But is this still valid, when we've had to replace $\sigma_{\bar{x}}$ with an estimate? Given a sample of size 36, it's close enough. Strictly speaking, the substitution of s for the unknown σ alters the shape of the sampling distribution. Instead of being Gaussian it now follows the t distribution, which looks very much like the Gaussian except that it's a bit "fatter in the tails".

5.1 *The Gaussian and t distributions*

Unlike the Gaussian, the t distribution is not fully characterized by its mean and standard deviation: there is an additional factor, namely the *degrees of freedom* (df). For the issue in question here—estimating a population mean—the df term is the sample size minus 1 (or 35, in the current example). At low degrees of freedom the t distribution is noticeably more "dispersed" than the Gaussian (for the same mean and standard deviation), which means that a 95 percent confidence would have to be wider, reflecting greater uncertainty. But as the degrees of freedom increase, the t distribution converges towards the Gaussian. By the time we've reached 30 degrees of freedom the two are almost indistinguishable. For the normal distribution, the values that enclose the central 95 percent are $\mu - 1.960\sigma$ and $\mu + 1.960\sigma$; for the t distribution with df = 30 the corresponding values are $\mu - 2.042\sigma$ and $\mu + 2.042\sigma$. Both are well approximated by the rule of thumb, $\mu \pm 2\sigma$.

5.2 *Further examples*

There's nothing sacred about 95 percent confidence. The following information regarding the Gaussian distribution enables you to construct a 99 percent confidence interval.

$$P(\mu - 2.58\sigma < x < \mu + 2.58\sigma) \approx 0.99$$

Thus the 99 percent interval is $\bar{x} \pm 2.58\sigma_{\bar{x}}$. If we want greater confidence that our interval straddles the unknown parameter value (99 percent versus 95 percent) then our interval must be wider (±2.58 standard errors versus ±2 standard errors).

Here's an example using a different statistic. An opinion polling agency questions a sample of 1200 people to assess the degree of support for candidate X. In the sample the proportion, p, indicating support for X is 56 percent or 0.56. Our single best guess at the population proportion, π, is then 0.56, but we can quantify our uncertainty over this figure. The standard error of p is $\sqrt{\pi(1-\pi)/n}$. The value of π is unknown but we can substitute p or, if we want to be conservative (i.e. ensure that we're not underestimating the width of the confidence interval), we can put $\pi = 0.5$, which maximizes the value of $\pi(1-\pi)$. On the latter procedure, the estimated standard error is $\sqrt{0.25/1200} = 0.0144$. The large sample justifies the Gaussian assumption for the sampling distribution, so our 95 percent confidence interval is

$$0.56 \pm 2 \times 0.0144 = 0.56 \pm 0.0289$$

This is the basis for the statement "accurate to within plus or minus 3 percent" that you often see attached to opinion poll results.

5.3 Generalizing the idea

The procedure outlined in this section is of very general application, so let me try to construct a more general statement of the principle. To avoid tying the exposition to any particular parameter, I'll use θ to denote a "generic parameter". The first step is to find an estimator (preferably an unbiased one) for θ, that is, a suitable statistic that we can calculate from sample data to yield an estimate, $\hat{\theta}$, of the parameter of interest; this value, our "single best guess" at θ, is called a *point estimate*. We now set a confidence level for our interval estimate; this is denoted generically by $1-\alpha$ (thus, for instance, the 95 percent confidence level corresponds to $\alpha = 0.05$). If the sampling distribution of $\hat{\theta}$ is symmetrical, we can express the interval estimate as

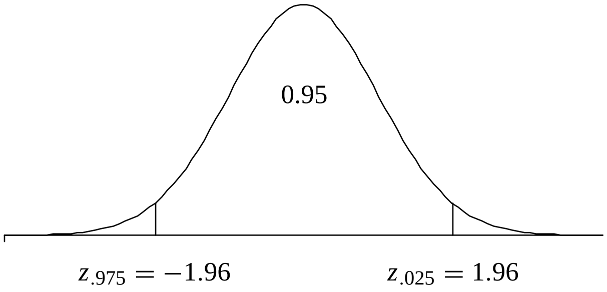
$$\hat{\theta} \pm \text{maximum error for } (1-\alpha) \text{ confidence}$$

The magnitude of the "maximum error" can be resolved into so many standard errors of such and such a size. The number of standard errors depends on the chosen confidence level (and also possibly on the degrees of freedom). The size of the standard error, $\sigma_{\hat{\theta}}$, depends on the nature of the parameter being estimated and the sample size.

Suppose the sampling distribution of $\hat{\theta}$ can be assumed to be Gaussian (which is often but not always the case). The following notation is useful:

$$z = \frac{x-\mu}{\sigma}$$

This "standard normal score" or "z-score" expresses the value of a variable in terms of its distance from the mean, measured in standard deviations. (Thus if $\mu = 1000$ and $\sigma = 50$, then the value $x = 850$ has a z-score of -3.0: it lies 3 standard deviations below the mean.) We can subscript z to indicate the proportion of the standard normal distribution that lies to its right. For instance, since the normal distribution is symmetrical, $z_{0.5} = 0$. It follows from points made earlier that $z_{0.025} = 1.96$ and $z_{0.005} = 2.58$. A picture may help to make this obvious.



Where the distribution of $\hat{\theta}$ is Gaussian, therefore, we can write the $1-\alpha$ confidence interval for θ as

$$\hat{\theta} \pm \sigma_{\hat{\theta}}\, z_{\alpha/2} \tag{7}$$

This is about as far as we can go in general terms. The specific formula for $\sigma_{\hat{\theta}}$ depends on the parameter.

Let me emphasize the last point, since people often seem to get it wrong. The standard error formula $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ may be the first one you encounter, but it is *not* universal: it applies only when we're using the sample mean to estimate a population mean. In general, each statistic has its own specific standard error. When a statistically savvy person encounters a new statistic, a common question would be, "What's its standard error?" *Warning*: it's not always possible to give an explicit formula in answer to this question (although it is for most of the statistics we'll come across in this course); in some cases standard errors have to be derived via computer simulations.

6 The logic of hypothesis testing

The interval estimation discussed above is a "non-committal" sort of statistical inference. We draw a sample, calculate a sample statistic, and use this to provide a point estimate of some parameter of interest along with a confidence interval. Often in econometrics we're interested in a more pointed sort of inference. We'd like to know whether or not some claim is consistent with the data. In other words, we want to *test hypotheses*.

There's a well-known and mostly apt analogy between the set-up of a hypothesis test and a court of law. The defendant on trial in the statistical court is the *null hypothesis*, some definite claim regarding a parameter of interest. Just as the defendant is presumed innocent until proved guilty, the null hypothesis is assumed true (at least for the sake of argument) until the evidence goes against it. The formal decision taken at the conclusion of a hypothesis test is either to *reject* the null hypothesis (cf. find the defendant guilty) or to *fail to reject* that hypothesis (cf. not guilty). The "fail to reject" locution may seem cumbersome (why not just say "accept"?) but there's a reason for it. Failing to reject a null hypothesis does *not* amount to proving that it's true. (Here the law court analogy falters, since a defendant who is found not guilty is entitled to claim innocence.)

The statistical decision is "reject" or "fail to reject". Meanwhile, the null hypothesis (often written H_0) is in fact either true or false. We can set up a matrix of possibilities.

	H_0 is in fact:	
Decision:	*True*	*False*
Reject	Type I error	Correct decision
Fail to reject	Correct decision	Type II error

Rejecting a true null hypothesis goes under the name of "Type I error". This is like a guilty verdict for a defendant who is really innocent. Failing to reject a false null hypothesis is called "Type II error": this corresponds to a guilty defendant being found not guilty. Since the hypothesis testing procedure is probabilistic, there is always some chance that one or other of these errors occurs. The probability of Type I error is labeled α and the probability of Type II error is labeled β. The quantity $1 - \beta$ has a name of its own: it is the "power" of a test. If β is the probability that a false null hypothesis will *not* be rejected, then $1 - \beta$ is the probability that a false hypothesis will indeed be rejected. It thus represents the power of a test to discriminate—to unmask false hypotheses, so to speak.

Obviously we would like for both α and β to be as small as possible. Unfortunately there's a trade-off. This is easily seen in the law court case. If we want to minimize the chance of innocent parties being found guilty, we can tighten up on regulations concerning police procedures, rules of evidence and so on. That's all very well, but inevitably it raises the chances that the courts will fail to secure guilty verdicts for some guilty parties (e.g. some people will get off on "technicalities").

The same issue arises in hypothesis testing, but in even more pointed form. We get to *choose in advance* the value of α, the probability of Type I error. This is also known as the "significance level" of the test. (And, yes, it's closely related to the α of confidence intervals, as we'll see before long.) While we want to choose a "small" value of α we're constrained by the fact that shrinking α is bound to crank up β, eroding the power of the test.

6.1 Choosing the significance level

How do we get to choose α? Here's a first approximation. The calculations that compose a hypothesis test are condensed in a key number, namely a conditional probability: *the probability of observing the given sample data,* *on the assumption that the null hypothesis is true*. If this probability, called the “p-value”, is small, we can place one of two interpretations on the situation: either (a) the null hypothesis is true and the sample we drew is an improbable, unrepresentative one, or (b) the null hypothesis is false (and the sample is not such an odd one). The smaller the p-value, the less comfortable we are with alternative (a). To reach a conclusion we must specify the limit of our comfort zone, or in other words a p-value below which we’ll reject H_0. Say we use a cutoff of .01: we’ll reject the null hypothesis if the p-value for the test is $\leq .01$. Suppose the null hypothesis is in fact true. What then is the probability of our rejecting it? It’s the probability of getting a p-value less than or equal to .01, which is (by definition) .01. In selecting our cutoff we selected α, the probability of Type I error.

If you’re thinking about this, there should be several questions in your mind at this point. But before developing the theoretical points further it may be useful to fix ideas by giving an example of a hypothesis test.

6.2 Example of hypothesis test

Suppose a maker of RAM chips claims an access time of 60 nanoseconds (ns) for the chips. The manufacture of computer memory is in part a probabilistic process; there’s no way the maker can guarantee that each chip meets the 60 ns spec. The claim must be that the average response time is 60 ns (and the variance is not too large). Quality control has the job of checking that the production process is maintaining acceptable access speed. To that end, they test a sample of chips each day. Today’s sample information is that with 100 chips tested, the mean access time is 63 ns with a standard deviation of 2 ns. Is this an acceptable result?

To put the question into the hypothesis testing framework, the first task is to formulate the hypotheses. Hypotheses, plural: we need both a null hypothesis and an alternative hypothesis (H_1) to run against H_0. One possibility would be to set $H_0\colon \mu = 60$ against $H_1\colon \mu \neq 60$. That would be a symmetrical setup, giving rise to a *two-tailed test*. But presumably we don’t mind if the memory chips are faster than advertised; we have a problem only if they’re slower. That suggests an asymmetrical setup, $H_0\colon \mu \leq 60$ (“the production process is OK”) versus $H_1\colon \mu > 60$ (“the process has a problem”).

We then need to select a significance level or α value for the test. Let’s go with .05.

The next step is to compute the p-value and compare it with the chosen α. The p-value, once again, is the probability of the observed sample data on the assumption that H_0 is true. The “observed sample data” will be summarized in a relevant statistic; since this test concerns a population mean, the relevant statistic is the sample mean. The p-value can be written as

$$P(\bar{x} \geq 63 \,|\, \mu \leq 60)$$

when $n = 100$ and $s = 2$. That is, if the population mean were really 60 or less, as stated by H_0, how probable is it that we would draw a sample of size 100 with the observed mean of 63 or greater, and a standard deviation of 2? Note the force of the “63 or greater”. With a continuous variable, the probability of drawing a sample with a mean of *exactly* 63 is effectively zero, regardless of the truth or falsity of the null hypothesis. We’re really asking, what are the chances of drawing a sample like this *or worse* (from the standpoint of the null hypothesis)?

We can assign a probability by using the sampling distribution concepts we discussed earlier. The sample mean (63) was drawn from a particular *distribution*, namely the sampling distribution of $\bar{x}$. If the null hypothesis is true, $E(\bar{x})$ is no greater than 60. The estimated standard error of $\bar{x}$ is $s/\sqrt{n} = 2/10 = .2$. With $n = 100$ we can take the sampling distribution to be normal. We use this information to formulate a *test statistic*, a statistic whose probability, on the assumption that H_0 is true, we can determine by reference to the standard tables. In this case (Gaussian sampling distribution) the test statistic is the z-score, introduced in section 5.3 above. In general terms, z equals “value minus mean, divided by standard deviation”. Here, the mean in question is the mean of the sampling distribution of $\bar{x}$, namely the population mean according to the null hypothesis or μ_{H_0}, while the relevant standard deviation is the standard error of $\bar{x}$. The The z-score formula is therefore

$$z = \frac{\bar{x} - \mu_{H_0}}{s_{\bar{x}}} = \frac{63 - 60}{.2} = 15$$

The p-value, therefore, equals the probability of drawing from a normal distribution a value that is 15 standard deviations above the mean. That is effectively zero: it’s far too small to be noted on any standard statistical tables.

At any rate it's much smaller than .05, so the decision must be to reject the null hypothesis. We are driven to the alternative, that the mean access time exceeds 60 ns and the production process has a problem.

6.3 Variations on the example

Suppose the test were as described above, except that the sample was of size 10 instead of 100. How would that alter the situation? Given the small sample and the fact that the population standard deviation, σ, is unknown, we could not justify the assumption of a Gaussian sampling distribution for $\bar{x}$. Rather, we'd have to use the t distribution with df = 9. The estimated standard error, $s_{\bar{x}} = 2/\sqrt{10} = 0.632$, and the test statistic is

$$t(9) = \frac{\bar{x} - \mu_{H_0}}{s_{\bar{x}}} = \frac{63 - 60}{.632} = 4.74$$

The p-value for this statistic is 0.000529—a lot larger than for $z = 15$, but still considerably smaller than the chosen significance level of 5 percent, so we still reject the null hypothesis.[1]

Note that, in general, the test statistic can be written as

$$\text{test} = \frac{\hat{\theta} - \theta_{H_0}}{s_{\hat{\theta}}}$$

That is, sample statistic minus the value stated in the null hypothesis—which by assumption equals $E(\hat{\theta})$—divided by the (estimated) standard error of $\hat{\theta}$. The distribution to which "test" must be referred, in order to obtain the p-value, depends on the situation.

Here's another variation. We chose an asymmetrical test setup above. What difference would it make if we went with the symmetrical version, $H_0\colon \mu = 60$ versus $H_1\colon \mu \neq 60$? This is the issue of one-tailed versus two-tailed tests. We have to think: *what sort of values of the test statistic should count against the null hypothesis?* In the asymmetrical case only values of $\bar{x}$ greater than 60 counted against H_0. A sample mean of (say) 57 would be quite consistent with $\mu \leq 60$; it is not even *prima facie* evidence against the null. Therefore the *critical region* of the sampling distribution (the region containing values that would cause us to reject the null) lies strictly in the upper tail. But if the null hypothesis were $\mu = 60$, then values of $\bar{x}$ both substantially below and substantially above 60 would count against it. The critical region would be divided into two portions, one in each tail of the sampling distribution. The practical consequence is that *we'd have to double the p-value found above*, before comparing it to α. The sample mean was 63, and the p-value was defined as the probability of drawing a sample "like this or worse", from the standpoint of H_0. In the symmetrical case, "like this or worse" means "with a sample mean this far away from the hypothesized population mean, or farther, in either direction". So the p-value is $P(\bar{x} \geq 63 \cup \bar{x} \leq 57)$, which is double the value we found previously. (As it happens, the p-values found above were so small that a doubling would not alter the result, namely rejection of H_0).

7 Hypothesis tests and p-values: further discussion

Let E denote the sample evidence and H denote the null hypothesis that is "on trial". The p-value can then be expressed as $P(E|H)$. This may seem an awkward formulation. Wouldn't it be better if we calculated the conditional probability the other way round, $P(H|E)$? Instead of working with the probability of obtaining a sample like the one we in fact obtained, assuming the null hypothesis to be true, why can't we think in terms of the probability that the null hypothesis is true, given the sample evidence we obtained? This would arguably be more "natural" and comprehensible.

To see what would be involved in the alternative approach, let's remind ourselves of the multiplication rule for probabilities, which we wrote as

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

[1] I determined the p-value using the econometric software package, `gretl`. I'll explain how to do this in class.

Swapping the positions of A and B we can equally well write

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

And taking these two equations together we can infer that

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

or

$$P(B|A) = \frac{P(B) \times P(A|B)}{P(A)} \tag{8}$$

The above equation is known as Bayes' rule, after the Rev. Thomas Bayes. It provides a means of converting from a conditional probability one way round to the inverse conditional probability. Substituting E for evidence and H for null hypothesis, we get

$$P(H|E) = \frac{P(H) \times P(E|H)}{P(E)}$$

We know how to find the p-value, $P(E|H)$. To obtain the probability we're now canvassing as an alternative, $P(H|E)$, we have to supply in addition $P(H)$ and $P(E)$. $P(H)$ is the marginal probability of the null hypothesis and $P(E)$ is the marginal probability of the sample evidence. Where are these going to come from?

7.1 Bayesian statistics

There is an approach to statistics that offers a route to supplying these probabilities and computing $P(H|E)$: it is known as the Bayesian approach, and it differs from the standard sampling theory doctrine. On the standard view, talking of $P(H)$ is problematic. The null hypothesis is in fact either true or false; it's not a probabilistic matter. Given a random sampling procedure, though, we *can* talk of a probability distribution for the sample statistic, and it's on this basis that we determine the p-value. Bayesians dispute this; they conceive probabilities in terms of *degree of justified belief* in propositions. Thus it's quite acceptable to talk of a $P(H)$ that differs from 0 or 1: yes, the hypothesis is *in fact* true or false, but we don't know which, and what matters is the degree of confidence we're justified in reposing in the hypothesis: this can be represented as a probability.

For a Bayesian, the $P(H)$ that appears on the right-hand side of Bayes' rule is conceived as a "prior probability". It's the degree of belief we have in H *before* seeing the evidence. The conditional probability on the left is the "posterior probability", the modified probability after seeing the sample. The rule provides an algorithm for modifying our probability judgments in the light of evidence.

One difficulty with the Bayesian approach is obtaining the prior probability. For instance, in the example above, it's not obvious how we should assign a probability to $\mu \leq 60$ in advance of seeing any sample data. There are techniques, however, for formulating "ignorance priors"—prior probabilities that correctly reflect an initial state of ignorance regarding the parameter values.

To illustrate the idea, let me vary the example above. Suppose the chip maker packages up RAM into boxes of one thousand modules, with a speed specification of either 60 ns or 70 ns. We're faced with a box whose label has come off: which sort does it contain? Suppose we set $H_0\colon \mu = 60$ against $H_1\colon \mu = 70$. If the 60 ns and 70 ns boxes are produced in equal numbers a suitable ignorance prior would be a $P(H)$ of 0.50 for the hypothesis that the mystery box contains 60 ns chips. We sample 9 of the chips and find a sample mean access time of 64 ns with a standard deviation of 3 ns. What then is the posterior probability of the hypothesis $\mu = 60$?

The standard test statistic is

$$t(8) = \frac{64 - 60}{3/\sqrt{9}} = 4.0$$

which has a two-tailed p-value of 0.004. At this point we have the prior, $P(H) = 0.50$, and the p-value, $(E|H_0) = 0.004$. What about the marginal probability of the evidence, $P(E)$? We have to decompose this as follows:

$$P(E) = P(E|H_0)P(H_0) + P(E|H_1)P(H_1)$$

which means we have another calculation to perform: $P(E|H_1)$. This is similar to the p-value calculation for H_0. We want the two-tailed p-value for

$$t(8) = \frac{64 - 70}{3/\sqrt{9}} = -6.0$$

which is 0.0003234. So:

$$P(H|E) = \frac{P(H) \times P(E|H)}{P(E)} = \frac{0.5 \times 0.004}{0.004 \times 0.5 + 0.0003234 \times 0.5} = 0.925$$

Based on the evidence, if the only two possibilities are that the sample chips came from a batch with a mean of 60 ns or a batch with a mean of 70, we can be fairly confident (92.5 percent) that they came from a 60 ns batch. Note that this seemed unlikely on the face of it (small p-value) but the probability of the evidence conditional on the alternative, $\mu = 70$, was much smaller still so the posterior probability of H_0 came out quite high. In this example $P(E|H_0) = .004$ yet $P(H_0|E) = .925$.

The Bayesian take on statistics is interesting and has quite a lot to recommend it, but in this course we'll concentrate on the standard sampling-theory approach. Thus you'll have to get used to thinking in terms of those "awkward" p-values! (Besides, as you've just seen, while the Bayesian approach does yield a value for the probability of the hypothesis conditional on the evidence it is not really a simplification; in fact it generally involves calculating the regular p-value and more. We need a prior probability for H_0 and the marginal probability of the sample, which are not required for the standard calculation.)

If you'd like to read more about Bayesian statistics here are two recommendations: *Data Analysis: A Bayesian Tutorial* by D. S. Sivia (Oxford: Clarendon Press, 1996), and the fascinating work by E. T. Jaynes, *Probability Theory: The Logic of Science*, online at `http://bayes.wustl.edu/etj/prob.html`.

8 Relationship between confidence interval and hypothesis test

We noted above that the symbol α is used for both the significance level of a hypothesis test (the probability of Type I error), and in denoting the confidence level $(1 - \alpha)$ for interval estimation. This is not coincidental. There is an equivalence between a two-tailed hypothesis test at significance level α and an interval estimate using confidence level $1 - \alpha$.

Suppose μ is unknown and a sample of size 64 yields $\bar{x} = 50$, $s = 10$. The 95 percent confidence interval for μ is then

$$50 \pm 1.96 \left(\frac{10}{\sqrt{64}} \right) = 50 \pm 2.45$$

Now suppose we want to test $H_0\colon \mu = 55$ using the 5 percent significance level. No additional calculation is needed. The value 55 lies outside of the 95 percent confidence interval, so we can immediately conclude that H_0 is rejected. In a two-tailed test at the 5 percent significance level, we fail to reject H_0 if and only if $\bar{x}$ falls within the central 95 percent of the sampling distribution, conditional on H_0, but since 55 exceeds 50 by more than the "maximum error", 2.45, we can see that, conversely, the central 95 percent of a sampling distribution centered on 55 will not include 50, so $\bar{x} = 50$ must lead to rejection of the null. "Significance level" and "confidence level" are complementary.