Audit Programs as Natural Experiments

Audit programs can best be thought of as stylized *natural experiments*. A natural experiment is an empirical study in which the activities of firms, individuals or groups are exposed to the experimental and control conditions that are determined by nature or by other factors outside the control of the auditors The process governing the exposures resembles 'random' assignment. The concept of 'randomness' is itself controversial, and I will briefly comment on this later in the chapter; it is intended to assure that samples are representative for some decision making objective. Natural experiments are observational studies and are *not* controlled in the traditional sense of an experiment. The difference between a natural experiment and a non-experimental observational study is that the former includes a comparison of conditions that pave the way for causal inference, but the latter does not. Natural experiments are employed as study designs when controlled experimentation is extremely difficult to implement or unethical, such as in several research areas addressed by epidemiology and economics. Field and quasi-experiments are closely related, but are not appropriate for audit programs; here are the differences:

1. Natural experiments rely on an external force (e.g. a government, nonprofit, etc.) controlling the randomization treatment assignment and implementation,

2. Field experiments require researchers to retain control over randomization and implementation.

3. Quasi-experiments occur when treatments are administered as-if randomly (e.g. U.S. Congressional districts where candidates win with slim-margins, weather patterns, natural disasters, etc.).

The perspective of audits as natural experiments motivates the methods applied to auditing in this book. It is also the reason that the R language is such an invaluable tool for auditing. At the core of this concept is the statistical analysis of *samples* to ultimately render the auditor's opinion.

Collecting and Analyzing Audit Evidence: Sampling

In statistical hypothesis testing (termed Neyman-Pearson hypothesis tests after the Neyman-Pearson Lemma), the *p-value* (probability value or asymptotic significance) is the probability for a given statistical model that, when the null hypothesis is true the sample mean would be greater than or equal to the actual observed results. Hypotheses are ways of bifurcating the decision space of a particular statistical inference task. In auditing, this bifurcation is a stylized *accept* or *do not accept* that a transaction stream is *in control* (interim tests) or an account balance is *fairly stated* or *is not fairly stated* (substantive tests).

Auditing makes decisions on control and fairness through searches for *material* or *intolerable errors* (the apportionment of *materiality* to individual accounts generates the set of *intolerable errors* for those accounts). Thus in an auditing context, the p-value of a test is the probability that the monetary error in the account balance would be *intolerable*.

One of auditing's idiosyncrasies is that audit tests consider only one-sided tests. Auditing, because of the *Conservatism Principal*, is concerned with overstating income and assets, or understating liabilities and expenses. As a consequence all audit tests are one-sided.

There are three types of sampling commonly used in audits:

1. *discovery sampling* for interim tests: *Discovery sampling* for interim tests sets an estimation sample size for *transaction-unit* samples, so that we are likely to discover at least one error in the sample if

the actual transaction error rate exceeds the *minimum acceptable error-rate* (alternatively called the *out-of-control* rate of error). Discovery tests helps the auditor decide whether the systems processing a particular transaction stream are in or out of control.

2. *attribute sampling* for interim tests: sets estimation sample size for *transaction-unit* samples to estimate the error rate in the entire transaction population with some confidence (e.g., 95%) that the estimate is within the *out-of-control* error-rate cutoff for that transaction stream. If it is found that a particular transaction stream is out of control, then attribute estimation will help us decide on the actual error rate of the systems that process this transaction stream. Errors estimates from attribute samples may either be *rates* or *amounts* or both.

3. *acceptance sampling* for substantive tests: Discovery sample analysis results in a decision of whether the internal control over a particular type of transaction error is "in control" or "out of control". If internal control is found to be insufficient, the auditor moves on to *attribute sampling* for interim tests, which sets estimation sample size for *transaction-unit* samples to estimate the error rate in the entire transaction population with some confidence (e.g., 95%) that the estimate is within the *out-of-control* error-rate cutoff for that transaction stream. If it is found that a particular transaction stream is out of control, then attribute estimation will help us decide on the actual error rate of the systems that process this transaction stream. Errors estimates from attribute samples may either be *rates* or *amounts* or both.

The sample sizes (an important consideration in determining the scope and budget of an audit) can be determined using Cohen's power analysis which is implemented in R's *pwr* package. Statistical 'power' is the complement (i.e., $1- \beta$) of the probability β of a type II error (type II error is the failure to reject a false null hypothesis, and is also known as a 'false negative.' Recall that, in statistical hypothesis testing, the probability α of a type I error is the 'significance' of the test (a type I error is the rejection of a true null hypothesis, and is also known as a 'false positive.'

To calculate the required sample size, you need to know four things:

- The size of the error to detect
- The variance of the response
- The desired significance level
- The desired power

```
x <- seq(-4, 4, length=100)
hx <- dnorm(x)
plot(x, hx, type="l", lwd=2,  col="red", xlim = c(-4,6), xlab="Error in
Account Balance",
  ylab="Density")
curve(dnorm(x, mean=3, sd=1),
      lwd=2, col="darkblue", add=TRUE, yaxt="n")
abline(v=1.7, col="gray80", lwd=3, lty=2)
text(3.7,.1, "Power", col = "gray20", adj = c(.8,.1))
text(3.3,-.01, "Significance", col = "gray20", adj = c(.8,.1))
text(1.4,.39, "Materiality", col = "gray20", adj = c(.4, -.1))
```

Sampling for Interim Tests of Compliance

Discovery sampling

Discovery sampling chooses a sample to determine whether an error rate does not exceed a designated percentage of the population. If the sample does not contain errors, then the actual error rate is assumed to be lower than the minimum unacceptable rate. The sampling calculation includes the following factors:

- Confidence level
- Minimum unacceptable error rate

Confidence level is a concept that originated in fiducial inference, an approach that has fallen out of fashion in favor of frequentist inference, Bayesian inference and decision theory. It is still used in decision theoretic approaches such as auditing, and roughly reflects the confidence that the auditor has in a particular decision (e.g., the balance is 'fairly stated'). Confidence is a concept that is intertwined with perceptions of 'risk' associated with audit failures. Some historical examples of failures are reviewed later in this chapter

Unacceptable error rates are parameters that are fixed at the start of an audit by the audit manager. They may be set by the firm, by prior years experience, or by some other method. In general, their choice is idiosyncratic determined by policies and perspectives of a particular firm or audit professional. Some of the differences in perspectives and audit focus of the 'Big Four' are discussed later in this chapter.

We can compute discovery sample size using what mathematicians call an *urn model*. Urn models are typically stated as draws of colored balls from an urn. In our case, we can consider the urn to be the set of all transactions of a given type that the firm processes in a given accounting period

Assume that the auditor determines that the *minimum acceptable error rate* for a particular transaction type (or alternately our *out-of-control* rate of error for that transaction type) is p. The discovery sample size needed for confidence c is $Pr[x \ge 0]=1-Pr[x=0]$ the probability that x, the number of errors discovered, is anything but 0

We can start by solving the probability of finding no errors in a sample of n draws from the urn:

$Pr[X \ge 0] = 1- {n \choose 0} \times p^0 \times (1-p)^{n-0} = 1-(1-p)^n$

For confidence level c we want to choose n so that:

$Pr[X \ge 0] = c \implies 1-(1-p)^n = c \implies n = \frac{log(1-c)}{log(1-p)}$

```
library(tidyverse)
library(reshape)

p <- seq(.0005,.015,.0005)
c <- .95
n <- log(1-c)/log(1-p)
samp <- data.frame(p,n)
ggplot(samp,aes(p,n))+
    geom_line()+
    labs(title="sample size for 95% confidence") +
```

```
      xlab("minimum acceptable error rate in population") +
      ylab("sample size")
```

Coefficient of variation formulas for sample size

If the auditor is not just concerned with the error rate, but also the stability of the error rate, then interim test sampling questions will involve both the error rate and the variability of that rate. If the *out-of-control* error rate is μ_1 and our goal is to detect an error of size Δ where the standard deviation of the sample is assumed to be σ then the sample size formula above becomes:

$n = \frac{8CV^2}{PE^2} [1+(1-PE)^2]$

Where PE is the proportionate error $PE = \frac{\Delta}{\mu_1}$ and CV is the coefficient of variation $CV=\frac{\sigma}{\mu_1}$. Sometimes the auditor will not have any idea of the variability of the population, but will still wish to consider it in sample size. In this case a variability of $\approx \frac{1}{3} \approx 35%$ is typical. In this case:

$n \approx \frac{1+(1-PE)^2}{PE^2} \approx \frac{2}{PE^2}$ for small error rates.

Rules of threes to calculate 95% upper confidence bounds

The rule of threes can be used to address the following type of question:

"The manager of the audit has told us that there have been no errors in the A/R account balance in the last 20 audits. Does this information give me an estimate whether there will be an error in this year's audit of A/R?"

The answer is "yes". Given no observed errors in the past n years, a $95%$ upper bound on the rate of occurrence is $\frac{3}{n}$

The basis of this is that, by the law of rare errors, the observed events Y (i.e., no errors in the audit) follow a $Poisson(\lambda)$ distribution using n samples. The sum of Poisson random variables is Poisson, so the question of at least one Y not equal to zero is the probability that the sum $\sum{Y_i}$ is greater than zero. Let this probability be, for example, 0.95 so that $P[\sum{Y_i} = 0] = e^{-n\lambda}=0.05$. Taking logarithms gives $n\lambda =-ln(0.05) \approx 3$. Thus the rate $\lambda \approx \frac{3}{n}$.

Attribute Sampling

Each time a control is tested, the auditor examines the evidence (samples, etc.) to make a decision on the hypothesis that the account is fairly stated, or internal control is effective. Since even an 'in-control' system will produce the occasional error, with an acceptable rate of error s. The auditor could make the assumption that $s=0$ (which would give a small sample size), but preferably a realistic estimate of actual error should be used. The Risk Assessment Matrix and review of prior years' audit papers for the control system will provide this rate s. The control decision takes the form of the following hypothesis test:

$H_0:$ The client's system is in control, controls are effective and error rate is approximately s.

$H_a:$ The client's controls are *not* effective, and the system is producing an intolerable number of errors $>r$.

Restating, if expected error rate is s, intolerable error rate is r and the error rate in the control system is ϵ then restate the hypotheses as:

$H_0: \epsilon=s$ where $\epsilon \in [0,1]$

$H_a: \epsilon \ge r$ where $r, \epsilon \in [0,1]$

This requires a one-sample proportion test to calculate power and sample size. The null hypothesis is $\epsilon=s$ and alternative hypothesis is $\epsilon>r$ with the standard power and significance assumptions of $\alpha = 0.05$ and $\beta = 0.8$. Let's test a $5% $ intolerable error rate $r=0.05$ and an expected error rate $s=0.01$:

```
library(pwr)
sample <- pwr.p.test(h = ES.h(p1 = 0.05, p2 = 0.01),
           sig.level = 0.05,
           power = 0.80,
           alternative = "greater")
sample
plot(sample)
```

Acceptance Sampling

The basic idea of calculating power or sample size with the *pwr* package is to leave out the argument that you want to calculate. If you want to calculate sample size, leave n out of the function.

To calculate power and sample size for one-sample t-tests, we need to set the type argument to "one.sample".

Each time an account is audited or a control is tested, the auditor examines the evidence (samples, etc.) to make a decision on the hypothesis that the account is fairly stated, or internal control is effective. The decision takes the form of a hypothesis test. For an account balance this is:

$H_0:$ The client's balance for the account is correct (error is zero)

$H_a:$ The client's balance for the account contained a material error.

If intolerable error (i.e., the part of materiality allocated to this account) is M and the error in the account balance is ϵ then restate this as:

$H_0: \epsilon=0$

$H_a: \epsilon \ge M$

In the alternative hypothesis, whether E and M are Dr or Cr is determined by the Conservatism Principle $-$ the direction that will produce the lowest income.

Let's start with a basic formula for distinguishing between a zero error μ_0 and the actual error μ_1 with a one-sided test, a Normal distribution probability model having homogeneous variances $\sigma_0^2 = \sigma_1^2 = \sigma^2$ and single sample compared to a known (i.e., $\mu_0 = 0$) distribution. The "rule of thumb" for sample size is:

$n = \frac{8}{\Delta^2}$

Where the error Δ we are trying top detect is:

$\Delta = \frac{\mu_1 - \mu_0}{\sigma}$

This is derived from the Normal sample formula:

$n = \frac{2(z_{1-\alpha}+z_{1-\beta})^2}{(\frac{\mu_1 - \mu_0}{\sigma})^2}$

For $\alpha=.05$, $\beta=.20$, $z_{1-\alpha}=1.96$ and $z_{1-\beta}=.84$ the value $2(z_{1-\alpha}+z_{1-\beta})^2 \approx 8$

This formula can also be used to calculate detectable error in the population using a given sample size n is $\Delta = \sqrt{\frac{8}{n}}$, the inversion of $n = \frac{8}{\Delta^2}$.

Suppose you are auditing a financial account, and want to compare the actual error in the account to an assumption (null hypothesis) that that account is fairly stated (has zero error. You will be measuring the actual expected error μ_2 with respect to an assumed error of μ_1 which will typically be zero. We can then define $\Delta = \mu_1 - \mu_2$. The smaller the difference you want to detect, the larger the required sample size.

Of the four variables that go into the sample size calculation, the variance of the responses can be the most difficult to determine. Usually, before you do your experiment, you don’t know what variance to expect. Investigators often conduct a pilot study to determine the expected variance, or information from a previous published study can be used.

The effect size combines the minimal relevant difference and the variability into one measurement $\frac{\Delta}{\sigma}$.

Significance α is often set at 0.05, for a 95% confidence, after a suggestion by RA Fisher in the 1920s.

$1- \beta$, where β is the probability of a Type 2 error (failing to reject the null hypothesis when the alternative hypothesis is true). In other words, if you have a 20% chance of failing to detect a real difference, then the power of your test is .8.

The calculation for the total sample size is:

n = $\frac{4 \times (z_{\alpha} + z_{\beta})^2 \times {\sigma}^2}{\Delta^2}$

```
library(pwr)

Delta <- 20
sigma <- 60
```

```
d <- Delta/sigma
sample <- pwr.t.test(d=d, sig.level=.05, power = .90, type = 'one.sample')
sample
plot(sample)
```

Acceptance Sampling with Poisson data

Substantive testing often takes a "monetary-unit" perspective of error, where each dollar in an account balance is assumed to be a sample-able unit. With monetary-units, data can be assumed to be Poisson distributed, taking on discrete dollar-unit values greater than zero (i.e., being left truncated at zero). The Poisson distribution has one parameter, the rate λ. The required sample size to detect an error rate of λ is:

$n \approx \frac{4}{\lambda}$

This is derived from the previous formula, noting that for a Poisson random variable Y, the transformed random variable $\sqrt{Y} \approx Normal({\mu=\sqrt{\lambda},\sigma^2=.25})$