Data Science and Statistical Computing Course Notes

Isaac Holt

January 9, 2023

1 Introduction

1.1 Standard errors

Theorem 1.1.1. (Central Limit Theorem) If we have a sample of data $(x_1, \ldots, x_n)$ where each $X_i \sim UK(\mu, \sigma^2)$ (UK is an unknown population distribution), then

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

as $n \to \infty$. This means that the distribution of the sample mean tends to a Normal distribution with standard deviation $\frac{\sigma}{\sqrt{n}}$. This is independent of UK.

Proof. Proven in Probability I. □

Definition 1.1.2. Let $\underline{x} = (x_1, \ldots, x_n)$ be a data set sampled from a population. The **unbiased estimate** of the **population standard deviation** is given by

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

Definition 1.1.3. The **standard error of the sample mean** is an estimate of the standard deviation of the sample mean and is given by

$$\widehat{\text{stddev}}(\bar{x}) = \frac{s}{\sqrt{n}}$$

where s is the unbiased estiate of the population standard deviation.

Remark. If n is not sufficiently large, then s is a poor estimator and the distribution of the sample mean might not be normally distributed.

If the population distribution is normal and n is small, then

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Definition 1.1.4. The equation above is called a **pivotal quantity**, because after transformation, the distribution no longer depends on the parameters (μ and σ) of the distribution of X.

1.2 Hypothesis tests

Definition 1.2.1. Given data $\underline{x} = (x_1, \ldots, x_n)$, we define a **null hypothesis**, H_0, to identify the distribution we believe generated each x_i.

Then we select a **test statistic**, $S(\cdot)$, which is a function of data which produces an extreme value when H_0 is false, and a value which is not extreme otherwise.

The **observed test statistic** for the data $\underline{x}$ is $t = S(x_1, \ldots, x_n)$. A hypothesis test compares the observed test statistic to the distribution of test statistic values that would occur assuming H_0 was true. This helps us decide whether the observed test statistic is extreme.

So we need to know the distribution of the random variable $T = S(X_1, \ldots, X_n)$, then we see how extreme t is as a realisation of T.

Definition 1.2.2. Given data $\underline{x} = (x_1, \ldots, x_n)$, let H_0 be a null hypopthesis that specifies a proposed distribution for the random variables X_i and let $S(\cdot)$ be a test statistic.

Let $t = S(x_1, \ldots, x_n)$ be the observed test statistic. If the distribution of $T = S(X_1, \ldots, X_n)$ is known, then the one-sided p-value is either

$$\mathbb{P}(T \geq t \mid H_0 \text{ true}) \quad \text{or} \quad \mathbb{P}(T \leq t \mid H_0 \text{ true})$$

If t is extreme when larger or smaller than T, the two-sided p-value is

$$\mathbb{P}(T \leq -|t| \cup T \geq |t| \mid H_0 \text{ true}) = \mathbb{P}(|T| \geq |t| \mid H_0 \text{ true})$$

2 Monte Carlo testing

2.1 Motivation for Monte Carlo testing

When conducting a hypothesis test, we have data $(x_1, \ldots, x_n)$ which follows a distribution $F(x \mid \theta)$ with parameter θ. We want to test

$$H_0 : \theta = \theta_0, \quad H_1 : \theta > \theta_0$$

Let $S(\cdot)$ be a test statistic, $T = S(X_1, \ldots, X_n)$, $t = S(x_1, \ldots, x_n)$ be the observed test statistic. To calculate the p-value, we use

$$\mathbb{P}(T \geq t \mid H_0 \text{ true}) = \mathbb{P}(T \geq t \mid X_i \sim F(\cdot, \theta_0))$$

This probability is often difficult or impossible to compute analytically.

But because $f(x \mid \theta)$ (the pdf of the distribution F) and S are known, we can estimate this probability using **simulation**.

2.2 Monte Carlo testing

Definition 2.2.1. Let $\underline{x}$ be observed data, $S(\cdot)$ be a test statistic, $t = S(\underline{x})$ be the observed test statistic, and $F(x \mid \theta)$ be a data-generating distribution. With hypotheses

$$H_0 : \theta = \theta_0, \quad H_1 : \theta > \theta_0$$

we can perform a **Monte Carlo hypothesis test** with the following algorithm:

For each $i \in \{1, \ldots, N\}$ (N is some large constant):

1. Simulate n observations $\underline{z} = (z_1, \ldots, z_n)$ from $Z_i \sim F(\cdot \mid \theta_0)$.
2. Compute $t_i = S(\underline{z})$.

Compute the estimated p-value with:

$$\mathbb{P}(T \geq t \mid H_0 \text{ true}) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{t_i \geq t\}$$

where

$$\mathbb{1}\{A\} = \begin{cases} 1 & \text{if } A \text{ true} \\ 0 & \text{if } A \text{ false} \end{cases}$$

is the indicator function.

Remark. This makes hypothesis testing much easier and allows it to be generalised to any distributions and test statistics.

1. We assume H_0 is true and simulate N sets of data (the $\underline{z}$'s) for each of which we compute the test statistic S.
2. We then count the number of times the test statistic of a simulated set of data was at least as extreme as the observed statistic t, then divide this total by the total number of simulated data sets, N.

Remark. Monte Carlo testing can only be performed when we know the true value of other parameters of F that are not being tested. For example, it can be done for Normal/t testing only when σ is known.

Example 2.2.2. Below is some R code which performs a Monte Carlo hypothesis test given some data, a test statistic function and a function that simulates random numbers from a distribution.

```
monte_carlo_p_value = function(sims, data, test_stat, rand) {
    # sims = N
    # data = x
    # test_stat = h
    # X = rand(...) <=> X ~ F(_, theta)
    obs_stat <- test_stat(data)
    sim_stats <- rep(0, sims)
    for (i in 1:N) {
        sim_data = rand(length(data))
        # obs = z
        sim_stat = test_stat(sim_data)
        # obs_stat = t_i
        sim_stats[i] <- sim_stat
    }

    return(sum(sim_stats > obs_stat) / sims)
}

# hypothesis test on mean candle lifetimes
# lifetimes are normally distributed
# H_0: mu = 9.2
# H_1: mu != 9.2
candle_lifetimes = c(8.1, 8.7, 9.2, 7.8, 8.4, 9.4)
mu0 = 9.2
N = 50000
lifetimes_stat = function(lifetimes) {
    return(abs(mean(lifetimes) - mu0))
}
gen_rand_lifetimes = function(sims) {
    return(rnorm(sims, mean = mu0, sd = sqrt(0.4)))
}

p_value = monte_carlo_p_value(N, candle_lifetimes, lifetimes_stat,
    gen_rand_lifetimes)
print(paste("p-value:", p_value))

```

3 The Bootstrap

3.1 Introduction

When doing Monte Carlo testing, we didn't need to know the distribution of the test statistic, but we did need to know the true population distribution in order to simulate data sets from that distribution. There is a method called the Bootstrap which allows us to estimate standard errors and confidence intervals without knowing the true population distribution and without knowing the test statistic distribution.

3.2 The non-parametric bootstrap

Definition 3.2.1. Let $\underline{x} = (x_1, \ldots, x_n)$ be a data set containing independent samples, let $S(\cdot)$ be a test statistic and let B be some large constant. To compute a **bootstrap estimate of the standard error** of S, draw B new samples **with replacement** (this is called **resampling**) of size n from $\underline{x}$ which gives $\underline{x}^{\star 1}, \ldots, \underline{x}^{\star B}$. Then calculate

$$\widehat{\mathrm{Var}}(S(\underline{x})) = \frac{1}{B-1} \sum_{i=1}^{B} \left(S(\underline{x}^{\star i}) - \bar{S}^{\star}\right)^2$$

where

$$\bar{S}^{\star} = \frac{1}{B} \sum_{i=1}^{B} S\left(\underline{x}^{\star i}\right)$$

The **bootstrap estimate of the statistic** is simply $S(\underline{x})$.

Remark. The Bootstrap is very powerful as the statistic S can be any statistic, e.g. mean, median, etc.

Example 3.2.2. Below is some R code which computes a Bootstrap estimate of a given statistic on a given data set.

```
1  bootstrap_estimate = function(samples, data, test_stat) {
2      # samples = B
3      # data = x
4      # test_stat = S
5      resamples <- rep(0, samples)
6      for (i in 1:samples) {
7          resampled <- sample(data, replace = TRUE)
8          resamples[i] <- test_stat(resampled)
9      }
10
11     return(sd(resamples))
12 }
13
14 # compute standard error of estimate of mean of data set
15 x <- c(94, 197, 16, 37, 99, 141, 23)
16 B <- 1000
17 S <- mean
18 estimate <- bootstrap_estimate(B, x, S)
19 print(paste("bootstrap estimate of standard error:", estimate))
20
```

3.3 Empirical distribution functions

Definition 3.3.1. For a random variable X that takes value in $\mathbb{R}$, the **distribution function** $F : \mathbb{R} \to [0, 1]$ (sometimes written F_X when dealing with multiple random variables) is defined as

$$F(x) := \mathbb{P}(X \leq x)$$

Remark. Defining a distribution function is enough to define the probability distribution of a random variable.

Definition 3.3.2. Let $(x_1, \ldots, x_n)$ be some data where each sample x_i is an i.i.d. (independent and identically distributed) realisation of a random variable X. The **empirical (cumulative) distribution function** (**ecdf**) is defined as

$$\hat{F}(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i \leq x\}$$

So the ecdf of x is the proportion of samples x_i which are less than or equal to x.

Proposition 3.3.3. The ecdf is a valid cdf.

Proof.

1. $\lim_{x \to -\infty} \hat{F}(x) = 0$ and $\lim_{x \to \infty} \hat{F}(x) = 1$
2. Monotonicity: for every $y < x, \hat{F}(y) \leq \hat{F}(x)$
3. Right-continuity: for every $x \in \mathbb{R}, \hat{F}(x) = \lim_{a \to x^+} \hat{F}(a)$, where $\lim_{a \to x^+} \hat{F}(a)$ is the limit from the right.

□

Theorem 3.3.4. (Glivenko-Cantelli) Let $X_1, \ldots, X_n$ be a random sample from a distribution with cdf $F(x)$. Then

$$\sup_{x \in \mathbb{R}} |\hat{F}(x) - F(x)| \to 0 \quad \text{as } n \to \infty$$

So as the number of random samples increases, the empirical cdf that they produce tends to the true cdf.

Proof. Too difficult. □

Example 3.3.5. Below is some R code which demonstrates the Glivenko-Cantelli theorem by plotting a graph containing the cdf of a given distribution and the ecdf generated by random samples from that distribution.

```
plot_ecdf_and_cdf = function(rand_samples, cdf) {
    # rand_samples = X_1, ..., X_n
    e <- ecdf(rand_samples)
    plot(e)
    points <- seq(-4, 4, length.out = 200)
    lines(points, cdf(points), lty = 2)
}

# plot ecdf and cdf of standard normal distribution
# ecdf uses 1000 samples
plot_ecdf_and_cdf(rnorm(1000), pnorm)

```

3.4 Sampling with replacement and ecdfs

Lemma 3.4.1. Let $\underline{x} = (x_1, \ldots, x_n)$ be a data set. Random uniform sampling from $\underline{x}$ is equivalent to sampling from the distribution that is defined by the ecdf $\hat{F}$ generated from $\underline{x}$.

Proof. We show that the cdf of the resampling distribution is the same function as the ecdf.

Sampling from replacement from $\underline{x}$ is equivalent to sampling from the discrete random variable Y which can take values in $\{x_1, \ldots, x_n\}$, where

$$p_Y(x_i) = \mathbb{P}(Y = x_i) = \frac{1}{n} \quad \forall i \in \{1, \ldots, n\}$$

and if $y \notin \underline{x}$ then $p_Y(y) = 0$.

From Probability I, the cdf for a probability mass function p is a piecewise constant:

$$\begin{aligned} F_Y(x) &= \sum_{t:t\leq x} p_Y(t) \\ &= \sum_{t:t\leq x} \frac{1}{n} \\ &= \frac{1}{n} \sum_{t:t\leq x} 1 \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i \leq x\} \\ &= \hat{F}(x) \end{aligned}$$

□

Remark. The ecdf approximates the true cdf and bootstrap resampling is equivalent to sampling from the ecdf, so is equivalent to fitting an ecdf to the data and then sampling from it as if it was a fitted distribution.

This is similar to Monte Carlo sampling, where we treat $\hat{F}$ as a known distribution.

Proposition 3.4.2. (Expectation and variance of the ecdf) Let $\underline{x} = (x_1, \ldots, x_n)$ be a data set and Y be a discrete random variable which can take values in $\{x_1, \ldots, x_n\}$ (so Y is defined by the ecdf generated by $\underline{x}$). Then $\mathbb{E}(Y) = \bar{x}$ and $\mathrm{Var}(Y) = \frac{n-1}{n} s_x^2$.

Proof.

$$\begin{aligned} \mathbb{E}(Y) &= \sum_{y\in\underline{x}} y p_Y(y) \\ &= \sum_{i=1}^{n} x_i \frac{1}{n} \\ &= \bar{x} \end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(Y) = &= \sum_{y \in \underline{x}} (y - \mathbb{E}(Y))^2 p_Y(y) \\
&= \sum_{i=1}^{n} (x_i - \bar{x})^2 \frac{1}{n} \\
&= \frac{n-1}{n} \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \frac{n-1}{n} s_x^2
\end{aligned}$$

□

Corollary 3.4.3. Let $\bar{Y}$ be the random variable that is equal to the mean of m independent samples, $Y_1, \ldots, Y_m$, from the ecdf generated by $(x_1, \ldots, x_n)$.

Then $\mathbb{E}(\bar{Y}) = \bar{x}$ and $\mathrm{Var}(\bar{Y}) = \frac{n-1}{n} \frac{s_x^2}{m}$.

Proof.

$$\begin{aligned}
\mathbb{E}(\bar{Y}) &= \mathbb{E}\left(\frac{1}{m} \sum_{i=1}^{m} Y_i \right) \\
&= \frac{1}{m} \mathbb{E}\left(\sum_{i=1}^{m} Y_i \right) \\
&= \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}(Y_i) \\
&= \frac{1}{m} \sum_{i=1}^{m} \bar{x} \\
&= \bar{x}
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(\bar{Y}) &= \mathrm{Var}\left(\frac{1}{m} \sum_{i=1}^{m} Y_i \right) \\
&= \frac{1}{m^2} \mathrm{Var}\left(\sum_{i=1}^{m} Y_i \right) \\
&= \frac{1}{m^2} \sum_{i=1}^{m} \mathrm{Var}(Y_i) \\
&= \frac{1}{m^2} \sum_{i=1}^{m} \frac{n-1}{n} s_x^2 \\
&= \frac{n-1}{n} \frac{s_x^2}{m}
\end{aligned}$$

□

Remark. The bootstrap standard error of a simple statistic, like the mean, can be calculated without running any simulations. E.g. if S is the mean, by the above corollary,

$$\widehat{\mathrm{Var}}(S(\underline{x})) \to \frac{n-1}{n} \frac{s^2}{n}$$

as the number of bootstrap samples tends to ∞, where s^2 is the sample variance.

3.5 The bootstrap with finite populations

Definition 3.5.1. Given a population of size N and a sample of size n. The **sampling fraction**, f is defined as

$$f = \frac{n}{N}$$

Remark. So far, we have been studying the **non-parametric bootstrap** which assumes a hypothetically infinite population size. In many cases, the population is not infinite but we can still assume it is if the population size, N, is much larger than the sample size, n. But generally if f is greater than roughly 0.1, we cannot make this assumption.

Let μ be the true population mean and σ^2 be the true population variance. As in the infinite population case, $\bar{X} = \frac{1}{n}\sum_{i=1}^{N} X_i$ is an unbiased estimate of μ, but unlike the infinite population case, the variance $\mathrm{Var}(\bar{X})$ is no longer $\frac{\sigma^2}{n}$, so the previous assumptions aren't valid.

Theorem 3.5.2. Given a finite population of size N, let $\bar{X}$ be the sample mean of a sample of size n, where each sample is drawn randomly and uniformly without replacement. Then

$$\mathrm{Var}(\bar{X}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$$

Proof. Let $c = \mathrm{Cov}(X_i, X_j) \neq 0 \quad \forall i \neq j$ (c does not change depending on i and j).

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathrm{Var}(X_i) + \sum_{i\neq j}\mathrm{Cov}(X_i, X_j)\right) \\
&= \frac{1}{n^2}\left(n\sigma^2 + \sum_{i\neq j} c\right) \\
&= \frac{1}{n^2}\left(n\sigma^2 + n(n-1)c\right) \\
&= \frac{1}{n}\left(\sigma^2 + (n-1)c\right)
\end{aligned}$$

To determine the value of c, consider the case where the whole population is sampled (this is possible as it is finite). Then $\bar{X} = \mu$ and $\mathrm{Var}(\bar{X}) = 0$, as there is no uncertainty in the mean. So

$$\mathrm{Var}(\bar{X}) = 0 = \frac{1}{N}(\sigma^2 + (N-1)c) \Longrightarrow c = \frac{-\sigma^2}{N-1}$$

and plugging this value of c into the expression for $\mathrm{Var}(\bar{X})$, we get

$$\mathrm{Var}(\bar{X}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$$

□

Remark. $\frac{N-n}{N-1} < 1$ so for a finite population, the standard error of the sample mean is less than in the infinite population case.

Remark. If N is large enough, $\frac{N-n}{N-1} \approx 1 - \frac{n}{N}$, therefore

$$\mathrm{Var}(\bar{X}) \approx (1-f)\frac{\sigma^2}{n}$$

so we see that a large sample size relative to the population size makes the difference between the finite and infinite population cases more significant.

Remark. $\lim_{N\to\infty} \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$ which is the standard error of the sample mean in the infinite population case, as expected.

Theorem 3.5.3. Given a finite population of size N and a sample of size n, where samples are drawn without replacement, let S^2 be the sample variance. Then an unbiased estimator of $\mathrm{Var}(\bar{X})$ is given by

$$\left(1 - \frac{n}{N}\right)\frac{S^2}{n}$$

Proof.

$$\begin{aligned}
\mathbb{E}(S^2) &= \frac{1}{n-1}\mathbb{E}\left(\sum_i \left(X_i - \bar{X}\right)^2\right) \\
&= \frac{1}{n-1}\mathbb{E}\left(\sum_i \left(X_i^2 - 2X_i\bar{X} + \bar{X}^2\right)\right) \\
&= \frac{1}{n-1}\mathbb{E}\left(\sum_i X_i^2\right) - \mathbb{E}\left(2\bar{X}\sum_i X_i\right) + \mathbb{E}\left(\sum_i \bar{X}^2\right) \\
&= \frac{1}{n-1}\left(\sum_i \mathbb{E}\left(X_i^2\right) - 2n\mathbb{E}\left(\bar{X}^2\right) + n\mathbb{E}\left(\bar{X}^2\right)\right) \\
&= \frac{n}{n-1}\left(\mathbb{E}\left(X_1^2\right) - \mathbb{E}\left(\bar{X}^2\right)\right)
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{E}\left(\left(1 - \frac{n}{N}\right)\frac{S^2}{n}\right) &= \frac{1}{n}\left(1 - \frac{n}{N}\right)\mathbb{E}(S^2) \\
&= \frac{1}{n-1}\left(1 - \frac{n}{N}\right)\left(\mathbb{E}\left(X_1^2\right) - \mathbb{E}\left(\bar{X}^2\right)\right)
\end{aligned}$$

Now, $\mathrm{Var}(X) = \mathbb{E}\left(X^2\right) - \mathbb{E}(X)^2$, therefore

$$\begin{aligned}
\mathbb{E}\left(\left(1 - \frac{n}{N}\right)\frac{S^2}{n}\right) &= \frac{1}{n-1}\left(1 - \frac{n}{N}\right)\left(\mathrm{Var}(X_1) + \mathbb{E}(X_1)^2 - \left(\mathrm{Var}(\bar{X}) + \mathbb{E}\left(\bar{X}\right)^2\right)\right) \\
&= \frac{1}{n-1}\left(1 - \frac{n}{N}\right)\left(\sigma^2 + \mu^2 - \left(\left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n} + \mu^2\right)\right)
\end{aligned}$$

which simplifies to

$$\mathbb{E}\left(\left(1 - \frac{n}{N}\right)\frac{S^2}{n}\right) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n} = \mathrm{Var}(\bar{X})$$

So it is an unbiased estimator. □

Remark. In the finite population case, the estimated bootstrap standard error of the mean tends to

$$\widehat{\mathrm{Var}}(S(\underline{x})) \to \frac{n-1}{n}\frac{s^2}{n}$$

and the true standard error of the mean is estimated by

$$\left(1-\frac{n}{N}\right)\frac{s^2}{n}$$

So the bootstrap estimate is incorrect (too big), unless $f = \frac{n}{N}$ satisfies $1-f = \frac{n-1}{n} \iff f = \frac{1}{n} \iff N = n^2$.

Remark. A solution to the problem described above is to increase the bootstrap resample size such that the standard error decreases until it equals the estimator of the true standard error.

By Corollary 3.4.3, if n' is the new sample size, then

$$\widehat{\mathrm{Var}}(S(\underline{x})) \to \frac{n-1}{n}\frac{s^2}{n'}$$

Now setting $n' = \frac{n-1}{1-f}$,

$$\widehat{\mathrm{Var}}(S(\underline{x})) \to \frac{n-1}{n}\frac{s^2}{n'} = (1-f)\frac{s^2}{n} = \left(1-\frac{n}{N}\right)\frac{s^2}{n}$$

Remark. The solution described above only works when we choose the test statistic to be the **mean**. For other choices of test statistic, we can use the method below:

Definition 3.5.4. Let $\underline{x} = (x_1, \ldots, x_n)$ be a data set containing independent samples taken from a finite population of size N. Let $S(\cdot)$ be the test statistic that we want to estimate and let B be a large constant. Assume that $k := \frac{N}{n}$ is an integer.

To perform the **population bootstrap**, first build a pseudo data set $\tilde{\underline{x}}$ by repeating $\underline{x}$ k times after itself, so $\tilde{x}$ has size N:

$$\tilde{\underline{x}} = (x_1, \ldots, x_n, x_1, x_n, \ldots, x_1, x_n)$$

Now perform the standard non-parametric bootstrap by taking B new samples of size n **without replacement** from $\tilde{\underline{x}}$ to give $\underline{x}^{\star 1}, \ldots, \underline{x}^{\star B}$. This is analogous to finite population sampling. Then as for the standard bootstrap,

$$\widehat{\mathrm{Var}}(S(\underline{x})) = \frac{1}{B-1}\sum_{i=1}^{B}\left(S\left(\underline{x}^{\star i}\right) - \bar{S}^{\star}\right)^2$$

where

$$\bar{S}^{\star} = \frac{1}{B}\sum_{i=1}^{B} S\left(\underline{x}^{\star i}\right)$$

More concisely,

$$\widehat{\mathrm{Var}}(S(\underline{x})) = \mathrm{Var}\left(\underline{x}^{\star 1}, \ldots, \underline{x}^{\star B}\right)$$

If $n \nmid N \iff N = kn + m$ for some $0 < m < n$, then before **each** bootstrap sample, build $\tilde{\underline{x}}$ by repeating $\underline{x}$ k times and append to $\tilde{\underline{x}}$ a sample of size m without replacement from $\underline{x}$.

3.6 The parametric bootstrap

If we have a strong belief that data follow a certain probability distribution, we can more accurately estimate the standard error using the **parametric bootstrap**, which involves estimating the distribution parameters and simulating data sets from it, instead of resampling the observed data.

Definition 3.6.1. Let $\underline{x} = (x_1, \ldots, x_n)$ be a data set containing independent samples that follow some distribution with parameter θ, $F(\cdot \mid \theta)$. Let $S(\cdot)$ be the test statistic and let B be a large constant. To construct a **parametric bootstrap estimate** of S and an estimate of the **standard error**:

1. Obtain the maximum likelihood estimator $\hat{\theta}$.
2. Take B samples of size n from $\hat{F}(\cdot) = F(\cdot \mid \hat{\theta})$ to give $\underline{x}^{\star 1}, \ldots, \underline{x}^{\star B}$. Then

$$\widehat{\mathrm{Var}}(S(\underline{x})) = \frac{1}{B-1} \sum_{i=1}^{B} \left(S\left(\underline{x}^{\star i}\right) - \bar{X}^{\star} \right)^2$$

   where

$$\bar{S}^{\star} = \frac{1}{B} \sum_{i=1}^{B} S\left(\underline{x}^{\star i}\right)$$

   This method is the same as the standard non-parametric bootstrap, apart from how samples are taken.

3.7 Bias estimation

Definition 3.7.1. When estimating a parameter θ of a distribution F using a statistic $S(\cdot)$ and observed data $\underline{x}$, for the estimate $\hat{\theta} = S(\underline{x})$, the **bias** is defined as

$$\mathrm{bias}(\theta, \hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta = \mathbb{E}(S(\underline{x})) - \theta$$

Remark. Sometimes a biased estimator can be preferable to an unbiased estimator if the biased estimator has lower variance and has a small bias.

Definition 3.7.2. Let $\underline{x}$ be a data set, $S(\cdot)$ be a test statistic, B be a large constant, θ be a distribution parameter and $\theta = S(\underline{x})$ be an estimator of θ. To calculate the **basic bootstrap bias estimate**, bootstrap resample from $\underline{x}$ to give $\underline{x}^{\star 1}, \ldots, \underline{x}^{\star B}$ and use

$$\widehat{\mathrm{bias}}(\theta, \hat{\theta}) = \bar{S}^{\star} - \hat{\theta} = \left(\frac{1}{B} \sum_{i=1}^{B} S\left(\underline{x}^{\star i}\right) \right) - S(\underline{x})$$

Remark. Notice the similarities with this formula and the definition of bias. Here, $\hat{\theta}$ is treated as the true value θ, and bootstrap resampling approximates $\mathbb{E}(\hat{\theta})$ with $\bar{S}^{\star}$.

Remark. Important: to apply a bias corrrection, subtract the bias from the estimate:

$$\hat{\theta} - \widehat{\mathrm{bias}}(\theta, \hat{\theta}) = \hat{\theta} - (\bar{S}^{\star} - \hat{\theta}) = 2\hat{\theta} - \bar{S}^{\star}$$

So the bias correct value is $2\hat{\theta} - \bar{S}^{\star}$, **not** $\bar{S}^{\star}$ itself.

Remark. Important: As explained in the previous remark, our estimate should be either $S(\underline{x})$ or $2\hat{\theta} - \bar{S}^{\star}$. But often $2\hat{\theta} - \bar{S}^{\star}$ has higher variance, so is not always preferable to $S(\underline{x})$. But calculating the level of bias is still useful, even if it is not used to adjust the estimate.

3.8 Confidence intervals

Definition 3.8.1. Statistics that we compute converge to a Normal distribution as the size of the sampled data increases, regardless of the distribution of the data.

In this case, the $100(1-\alpha)\%$ **normal confidence interval** is

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(S(\underline{x}))}$$

where $z_{\alpha/2}$ is the $100(\alpha/2)\%$ percentile of the standard Normal distribution.

Remark. We should check that the statistic roughly follows a Normal distribution before using this, especially if the sample size is small.

Remark. Important: here, we want to check Normality of the **bootstrap samples**, not the data itself, because the bootstrap samples indicate the sampling distribution of the statistic, and we want normality of the **statistic**, not the data.

Remark. Important: the confidence interval is centred on $\hat{\theta} = S(\underline{x})$, not $\bar{S}^{\star}$.

4 Monte Carlo Integration

Assume the integral we want to evaluate can be written as an expectation, with respect to some random variable Y, taking values in Ω, with pdf $f_Y(\cdot)$.

Then $\mu := \mathbb{E}[Y] = \int_\Omega y f_Y(y) dy$

If we assume μ exists, then we approximate μ with

$$\mu \approx \hat{\mu}_n := \frac{1}{n} \sum_{i=1}^{n} Y_i$$

where $Y_1, \ldots, Y_n$ are i.i.d. simulations from the distribution of Y.

By the weak law of large numbers,

$$\lim_{n \to \infty} \mathbb{P}(|\hat{\mu}_n - \mu| \leq \epsilon) = 1$$

Remark. $\hat{\mu}_n$ itself is a random variable

Often we can write $Y = g(X)$ for some random variable X with pdf $f_X(x)$. Then

$$\mu = \mathbb{E}[Y] = \mathbb{E}[g(X)] := \frac{1}{n} \sum_{i=1}^{n} g(X_i)$$

4.1 Accuracy

Assume $Var(Y) = \sigma^2 < \infty$, then

$$\mathbb{E}[\hat{\mu}_n] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Y_i] = \mu$$

and

$$Var(\hat{\mu}_n) = \mathbb{E}[(\hat{\mu}_n - \mu)^2] = \frac{\sigma^2}{n}$$

$\frac{\sigma^2}{n}$ is the mean square error (MSE).

$\frac{\sigma}{\sqrt{n}}$ is the root mean square error (RMSE).

To improve accuracy, we can control $\sqrt{n}$ (i.e. increase number of simulations). So we say RMSE is $O(n^{-1/2})$.

Definition 4.1.1. For functions f and g, $f(n) = O(g(n))$ if for some $C \in \mathbb{R}$, $n_0 \in \mathbb{R}$, $|f(n)| \leq Cg(n)$ for every $n \geq n_0$.

So for example, to reduce the error by half, we must increase number of simulations by factor of 4.

An extra decimal place of accuracy requires 100 times as many simulations.

This quickly grows to infeasible numbers of simulations.

4.2 Error estimation

Options:

Apply Chebyshev:

$$\mathbb{P}(|\hat{\mu}_n - \mu| \geq \epsilon) \leq \frac{\mathbb{E}((\hat{\mu}_n - \mu|)^2}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

OR use i.i.d Central Limit Theorem:

$$\lim_{n\to\infty} \mathbb{P}(\frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}} \leq z) = \Phi(z)$$

NB: we choose the value of n.

So select large enough n to be confident that the CLT applies.

i.e. we form a $100(1-\alpha)\%$ confidence interval $\hat{\mu}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

There is an important special case where $g(.)$ is some constant multiple of an indicator:

$$g(x) = c\mathbb{I}\{A(x)\}$$

We are estimating a probability here: $\hat{p}_n = \frac{1}{n}\sum_{i=1} n\mathbb{I}\{A(x)\}$ so confidence interval is $c\hat{p}_n \pm cz_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$

Problem 1: if no 1's are observed, then the CI (confidence interval) is $[0, 0]$.

Probability of getting no 1's in n simulations is $(1-p)^n$

So create CI by looking for maximal p such that this probability is at least α, i.e. $p \leq 1 - \alpha^{1/n} \approx -\frac{\log\alpha}{n}$.

Problem 2: very few number of simulations are 1.

If $\hat{p}$ is close to zero, $cz_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \approx cz_{\alpha/2}\sqrt{\frac{\hat{p}_n}{n}}$

Relative error is $cz_{\alpha/2}\sqrt{\frac{\hat{p}_n}{n}}\ \hat{p}_n = \frac{cz_{\alpha/2}}{\sqrt{\hat{p}_n n}}$

If we want the relative error to be at most δ then $n \geq \frac{c^2 {z_{\alpha/2}}^2}{\delta^2 \hat{p}_n}$. This grows very quickly when the event has a very low probability of occurring.

4.3 Notes on generality of expectation

- f_Y can be any valid pdf.
- Expectations are a very general tool. We can write any probability $\mathbb{P}(X < a)$ as an expectation $\mathbb{E}(\mathbb{I}\{X \in [-\infty, a]\})$, and the general case $\mathbb{P}(X \in E) = \mathbb{E}(\mathbb{I}\{X \in E\})$

4.4 Simulation

Ongoing assumption: we have access to a stream of uniform random numbers:

$$u_1, \ldots, u_n \sim Unif(0, 1)$$

Inverse Transform Sampler:

We want to simulate from a distribution $F(\cdot)$ (a cdf).

If F has an inverse F^{-1}, then to perform inverse transform sampling:

1. Simulate $U \sim Unif(0, 1)$

2. Compute $X = F^{-1}(u)$

Then $X \sim F(\cdot)$
A cdf F is valid if:

1. $\lim_{n\to-\infty} F(x) = 0$ and $\lim_{n\to\infty} F(x) = 1$
2. Monotonicity: $x' < x \Rightarrow F(x') \leq F(x)$
3. Right continuity: for every $x \in \mathbb{R}$, $F(x) = F(x^+)$ where $F(x^+)$ is the limit from the right.

Definition 4.4.1. Let F be a valid cdf. The generalised inverse cdf is $F^{-1}(u) = \inf\{x : F(x) \geq u\}$ for every $u \in [0, 1]$.

Theorem 4.4.2. Let F be a cdf with the generalised inverse cdf F^{-1}. If $U \sim \text{Unif}(0, 1)$ and $X = F^{-1}(U)$, $X \sim F$.

Proof. The cdf completely defines the distribution of X. By definition, $F^{-1}(U) \leq x \Leftrightarrow U \leq F(x)$, therefore $\mathbb{P}(X < x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F_U(F_X(x)) = F(x)$ (because $0 < F(x) < 1$ for every $x \in \mathbb{R}$). □

4.5 Rejection sampling

Rejection sampling allows the user of densities instead of distributinos, i.e. pdf instead of cdf.

Idea: to simulate from f, the target pdf, we instead simulate from another density close to it, called $\tilde{f}$, discarding away exactly the right number of simulations to be left with simulations from f. To do this, we need $f(x) \leq c\tilde{f}(x)$ for every $x \in \mathbb{R}^d$ where $c \leq \infty$. f is the target pdf, $\tilde{f}$ is the proposal pdf.

Definition 4.5.1. Given f and $\tilde{f}$ such that $f(x) \leq c\tilde{f}(x)$, we generate a rejection sample by:

1. $a =$ false
2. while a is false:
   (a) $U \sim \text{Unif}(0, 1)$
   (b) $X \sim \tilde{f}$
   (c) If $u \leq \frac{f(x)}{c\tilde{f}(x)}$ then set $a =$ true.
3. return X as a sample from f.

Lemma 4.5.2. The expected number of iterations required before the proposal is accepted is c.

Proof. Let A be the random variable indicating acceptance of the proposal.

$$\mathbb{P}(A = 1) = \int_\Omega \mathbb{P}(A = 1|X = x)\mathbb{P}(X = x)dx = \int_\Omega \mathbb{P}(u \leq \frac{f(x)}{c\tilde{f}(x)})\tilde{f}(x)dx$$

$$= \int_\Omega \frac{f(x)}{c\tilde{f}(x)} \tilde{f}(x)dx = \frac{1}{c} \cdot 1 = \frac{1}{c}$$

Therefore the number of iterations to acceptance is geometrically distributed, so

$$\mathbb{E}(\text{number of iterations to accept}) = 1/p = \frac{1}{1/c} = c$$

□

Theorem 4.5.3. Let $f, \tilde{f}$ be pdf's such that $f(x) < c\tilde{f}(x) \forall x \in \mathbb{R}^d$ for some $c < \infty$. Then X generated by rejection sampling is distributed as $f(\cdot)$.

Proof. Let $E \subset \Omega$.

$$\mathbb{P}(X \in E | A = 1) = \frac{\mathbb{P}(A = 1 | X \in E)\mathbb{P}(X \in E)}{\mathbb{P}(A = 1)} = \frac{\int_E \frac{f(x)}{c\tilde{f}(x)dx} \tilde{f}(x)}{1/c} = \int_E f(x)dx$$

which is $\mathbb{P}$(event E under pdf). □

Remark. We cannot always find a suitable c for all pairs $f, \tilde{f}$.

Remark. When calculating value of c, always round up if rounding is necessary.

4.6 Importance Sampling

Definition 4.6.1. Given a normalised pdf f and a normalised pdf $\tilde{f}$, we produce n **importance sample** by, for $i \in \{0, \dots, n\}$:

1. Generate $x_i \sim \tilde{f}(\cdot)$ - this could be with inverse transform or rejection sampling.
2. Compute $w_i = \frac{f(x_i)}{\tilde{f}(x_i)}$

$\{(x_i, w_i)\}_{i=1}^n$ are the importance samples.
We estimate $\mu = \mathbb{E}(g(X)) \approx \hat{\mu} = \frac{1}{n} \sum_{i=1}^n w_i g(x_i)$ where $X_i \sim \tilde{f}(\cdot)$.

$$\mathbb{E}(Xh(X)) = \int (xh(x))f(x)dx$$

If $h(x)f(x) =: \eta(x)$ is a valid pdf, then

$$\mathbb{E}_f(Xh(x)) = \int (xh(x))f(x)dx = \int xh(x)f(x)dx = \int x\eta(x)dx = \mathbb{E}_\eta(x)$$

Theorem 4.6.2. Let $\mu = \mathbb{E}_f(g(X))$ and let $\tilde{f}$ be a pdf such that if $g(x)f(x) \neq 0$, $\tilde{f}(x) > 0$.

Then $\hat{\mu}$ as defined in the definition satisfies:

$$\mathbb{E}_{\tilde{f}}(\hat{\mu}) = \mu$$

Proof.

$$\mathbb{E}_f(g(X)) = \int_\Omega g(x)f(x)dx = \int_\Omega g(x)\frac{\tilde{f}(x)}{\tilde{f}(x)}f(x)dx$$

$$= \left(g(x)\frac{f(x)}{\tilde{f}(x)}\right)\tilde{f}(x)dx = \mathbb{E}_{\tilde{f}}(\frac{g(X)f(X)}{\tilde{f}(X)}) = \mathbb{E}_{\tilde{f}}(w(X)g(X)) = \mathbb{E}_{\tilde{f}}(\hat{\mu})$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} w_i g(x_i) \to \mathbb{E}_{\tilde{f}}(w(X)g(X)) = \mathbb{E}_f(g(X))$$

as $n \to \infty$. □

Theorem 4.6.3. The variance of the importance sampled estimator is

$$\mathrm{Var}(\hat{\mu}) = \frac{\sigma^2_{\tilde{f}}}{n}$$

where

$$\sigma^2_{\tilde{f}} = \int_{\hat{\Omega}} \frac{(g(x)f(x) - \mu\tilde{f}(x))^2}{\tilde{f}(x)}dx$$

The optimal proposal to minimise $\sigma^2_{\tilde{f}}$ is

$$\tilde{f}_{\mathrm{opt}}(x) = \frac{|g(x)|f(x)}{\int |g(x)|f(x)dx}$$

Remark. Importance sampling can completely fail, so diagnostics are important here.

4.7 Self-normalised importance sampling

Definition 4.7.1. If f and/or $\tilde{f}$ are unnormalised pdfs, we modify the estimator:

$$\hat{\mu} = \frac{\sum_{i=1}^n w_i g(x_i)}{\sum_{i=1}^n w_i}$$

Remark.

1. $\hat{\mu}$ is **not** unbiased for finite simulations.
2. The variance has only the approximate form:

$$\mathrm{Var}(\hat{\mu}) \approx \frac{\hat{\sigma}^2_{\tilde{f}}}{n}$$

   where $\hat{\sigma}^2_{\tilde{f}} = \sum_{j=1}^n w_j'^2(g(x_i) - \hat{\mu})^2$ and $w_j' = \frac{w_j}{\sum_{i=1}^n w_i}$
3. The theoretically optimal proposal pdf is now

$$\tilde{f}_{\mathrm{opt}} \propto |g(x) - \mu|f(x)$$

Remark. A common way importance sampling can perform poorly is when there are a few simulations with large weights, which leads to a high variance of the estimator.

Options for diagnostics: ask what simulation size would give the same variance if we had done standard i.i.d. sampling from f, i.e. if the simulation variance of the expectation quanityt is σ^2 and we had n_e i.i.d. simulations from f, then $\text{Var}(\hat{\mu}) = \frac{\sigma^2}{n_e}$ which gives that

$$n_e = \frac{\left(\sum w_i\right)^2}{\sum w_i^2} = \frac{n\bar{w}^2}{\overline{w^2}}$$

where $\overline{w^2} = \frac{1}{n}\sum w_i^2$ and $\bar{w}^2 = \left(\frac{1}{n}\sum w_i\right)^2$

Remark. A low n_e is desirable, a high n_e is not desirable.

Definition 4.7.2. If we have bootstrap samples $S(\underline{x}^{\star b})$, $b \in \{1, \ldots, B\}$, then we can **order statistic notation** to show the ith largest value as $S^{\star}_{(i)}$. Then

$$S^{\star}_{(1)} \leq S^{\star}_{(2)} \leq \cdots \leq S^{\star}_{(B)}$$

Definition 4.7.3. If $\hat{F}(\cdot) = F(\cdot \mid \hat{\theta})$ is close to the true distribution, then the bootstrap samples $S(\underline{x}^{\star b})$, $b \in \{1, \ldots, B\}$ will hopefully follow the sampling distributino of the statistic (but centred on $\hat{\theta}$) instead). Then $100(1-\alpha)\%$ confidence interval using the **percentile CI method** is

$$\left[S^{\star}_{((\alpha/2)B)}, S^{\star}_{((1-\alpha/2)B)}\right]$$

B must be chosen so that $(\alpha/2)B$ and $(1-\alpha/2)B$ are integers.

Remark. It is important here that B is large. Often $B > 2000$ is required for a reasonable estimate using the percentile CI method.

Remark. This method doesn't require us to assume that the statistic is normally distributed, but is inaccurate if there is any bias or non-constant standard error. In the case where there is bias or a non-constant standard error, two advanced methods, BC (bias corrected) and BCa (bias corrected and accelerated) CIs, can be used.