Data Science and Statistical Computing Course Notes

Isaac Holt

December 30, 2022

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.

Definition 1.1.2. The unbiased estimate of the population standard deviation is given by

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

Then the **standard error of the sample mean** is $\frac{s}{\sqrt{n}}$, which is an estimate of the standard deviation of the sample mean.

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 1.1.3. 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**, $t = h(x_1, \ldots, x_n)$ 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 $t = h(x_1, \ldots, x_n)$. 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 $T = h(X_1, \ldots, X_n)$, 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 h be a test statistic.

Let $t = h(x_1, \ldots, x_n)$ be the observed test statistic. If the distribution of $T = h(X_1, \ldots, X_n)$ 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 $F(x|\theta)$ with parameter θ. We want to test

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

We define the test statistic as $T = h(X_1, \ldots, X_n)$ with an observed value of $t = h(x_1, \ldots, x_n)$. 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 h are known, we can estimate this probability using **simulation**.

2.2 Monte Carlo testing

Definition 2.2.1. Let $\underline{x}$ be observed data, $t = h(\underline{x})$ be an observed test statistic, $F(x \mid \theta)$ 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 $t_i = h(\underline{z})$.

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 $(\underline{z})$ for each of which we compute the test statistic h.
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

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 $f_Y(.)$.

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 $\mathrm{Var}(\hat{\mu}) = \frac{\sigma^2}{n_e}$ 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.