Big Data & Predictive Analytics
Lecture Notes on Simulation & Hypothesis Testing

School of Informatics
University of Leicester

1 Simulation of Distributions for Random Variables

What if you have a distribution that is so complex that you cannot write down its probability density function using a formula? In that case you can use a computer code to represent your distribution. This computer code is likely to rely upon random numbers. This process is called *simulation*. The simulation should enables us to encode the distribution of a complex random variable X and to calculate probabilities $P(X \leq x)$.

1.1 A motivating example

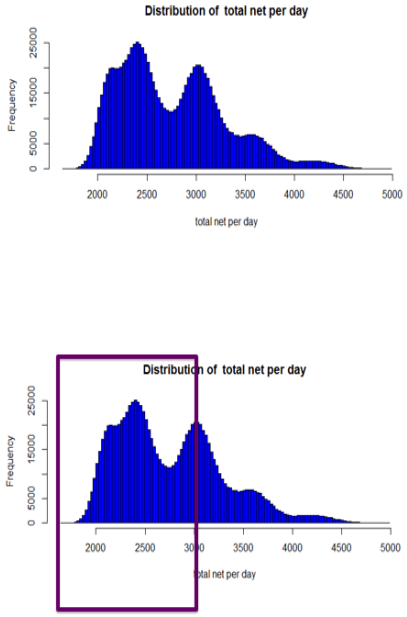
Consider a 'drink shop' or a lemonade stand. The profitability of the lemonade stand depends on the number of customers arriving, the weather since in sunny days we sell more drinks, the profit from the drinks customers order, and the tips the customer may or may not choose to leave. The distribution of possible profits is thus, the joint distribution of customer arrivals, items ordered, and tips. In practice, such a complex distribution cannot be analyzed except using simulation.

We are interested to find out the probability $P(X \leq x)$ that the profit will fall below a particular value x. In order to answer this question, we need to know exactly how the profit is calculated.

$$Profit = NumberCustomers \times (ProfitPerCup + Tip)$$

It is the number of customers, times how much profit we make per cup, and then plus any tips that the customers want to give us.

To model the profit, we need to model each variable in it. We can assume that the number of customers follows a normal distribution, say with mean 600 and standard deviation 30. Why did we choose normal? Because customers are independent Bernoulli's and thanks to the Central Limit Theorem, the sum of those random variables looks normal for a large enough sample size.



Now for the profit per cup, that depends on the weather. Because if the weather is hot, we can charge \$5 for a lemonade. If its medium, we charge \$4. But if the weather is bad, we can only justify charging \$3.50.

As for the tips, the customers give us whatever they want, but usually it's easiest for them to give us whatever change they have, so they might give us nothing, or a quarter, or one dollar or two dollars. Half the time they don't give us any tip. Which is OK considering how much we are charging for the lemonade.

1.2 Simulation in Python

The simulation is carried out daily. For each day, we simulate the number of customers from a normal distribution centered at 600 with std 30. Then we simulate the profit per cup as a discrete random variable depending on the weather that day. Then for each customer, we will simulate the tip they are going to give. Using this code, we can simulate as many days as we like. Then we can use the simulation results to actually create a histogram to show what the distribution of the profits are going to look like. We obtain the following distribution for the profit:

The bumps in the above distribution can be explained by the distributions of the profit per cup and the tip. Looking at each of those distribution, we can see how bumps get formed with various effects.

Back to the question of how to compute $P(X \leq x)$. If we want to know how often we will get less than \$3000 profit, we just need to count the percent of days when the profit is less than 3000.

In short, to simulate a complex distribution we can rely upon basic distributions such as uniform, normal, binomial, Bernoulli or Poisson. But we can also simulate piece wise defined

functions $y = f(x)$ by using say the uniform distribution in the interval $[0, 1]$ as follows:

$$\begin{aligned} y = 0 \quad if \quad x \leq 0.5 \\ y = 3.5 \quad if \quad 0.5 < x \leq 0.7 \\ y = 4 \quad if \quad 0.7 < x \leq 0.9 \\ y = 5 \quad if \quad 0.9 < x \leq 1.0 \end{aligned}$$

2 Confidence Intervals

2.1 Sampling error

Conceptually, the sampling error represents the variation due to selecting a single sample from a population. For example, the plus-or-minus margin stated in poll results represents a sampling error. In practice, we use only one sample to estimate the population parameter. To account for the variations due to using different samples, we estimate a *confidence interval*.

2.2 Estimating a confidence interval

A confidence interval represents a range of values delimited by a lower bound and an upper bound stated with a specific degree of certainty for some population parameter. For example, 'an interval estimate with 95% confidence' means that if all possible samples of the sample size were used in lieu of our single sample, 95% of the interval estimates obtained would include the population parameter but 5% would not. In essence, high levels of confidence lead to wider intervals. In general, 95% confidence intervals are considered an acceptable trade-off even though a 99% confidence interval (resulting in a wider interval) and 90% confidence interval (resulting a narrower interval) can be used in some cases. The confidence interval $[c_1, c_2]$ for the mean μ is such that the probability

$$P(c_1 \leq \mu \leq c_2) = 1 - \alpha$$

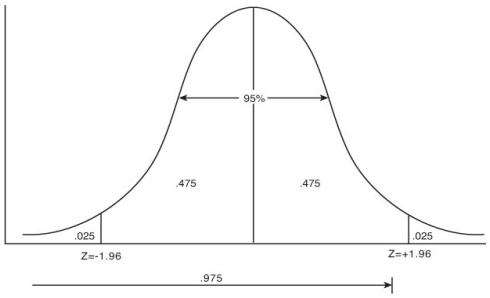
wherein α is the significance level e.g., $0.05, 0.10, 0.01$. Note that $100(1 - \alpha)$ gives us the confidence level e.g., $95\%, 90\%, 99\%$.

Estimating a confidence interval for the mean requires the knowledge of the sample mean but also the population standard deviation. However, the population's standard deviation is rarely known. We will consider the following three cases wherein the first two cases concern the estimation of the mean for a numerical variable and the third case concerns proportions for a categorical variable:

2.2.1 The population standard deviation is known

The Normal distribution tells that if we randomly select a member of a normally distributed population $N(\mu, \sigma)$, there is a 68% probability that the selected member lies in the interval $[\mu - \sigma, \mu + \sigma]$. Relying upon the central limit theorem, the variable

$$Z = \frac{X - \mu}{(\sigma/\sqrt{n})}$$



for a given variable X follows the normal distribution $N(0,1)$ under some conditions. We can then estimate the confidence interval as the range of values in the interval

$$\left[\mu - z\frac{\sigma}{\sqrt{n}}, \mu + z\frac{\sigma}{\sqrt{n}}\right]$$

in which n is the size of the sample and z is the *critical value* from the normal distribution.

The value z is obtained from the normal distribution table or by using a software package such as Python. Naturally, we need the inverse function of the cumulative distribution.

```
>>> import scipy.stats
>>> scipy.stats.norm(0,1).ppf(0.025)
-1.9599639845400545
>>> scipy.stats.norm(0,1).ppf(0.975)
1.959963984540054
```

2.3 The population standard deviation is unknown

If the population standard deviation is unknown, we rely upon the standard deviation S of the sample. The normalisation

$$Z = \frac{X - \mu}{(S/\sqrt{n})}$$

follows a *t-distribution* (similar to the normal distribution but with higher variance) with $n-1$ degrees of freedom. The confidence interval is given as

$$\left[\mu - t_{n-1}\frac{S}{\sqrt{n}}, \mu + t_{n-1}\frac{S}{\sqrt{n}}\right]$$

The t-value t_{n-1} is obtained using a t-distribution table or a software package such as Python. A fragment of the Python code we could write assuming we have calculated the sample mean `sm`, the sample standard deviation `ss`, the sample size `n` and the expected population mean `m`.

```
>>> from scipy import stats
>>> from scipy.stats import norm
>>> import numpy as np
>>> tv = (sm-m)/(ss/np.sqrt(n))  # t-statistic for mean
>>> pval = stats.t.sf(np.abs(tv), n-1)*2  # two-sided pvalue = Prob(abs(t)>tv)
```

2.3.1 The variable is categorical

If we are interested in a proportion p for a categorical variable, then the distribution tends to a normal distribution as the sample size increases. The confidence interval for the proportion

p is then estimated as

$$\left[p - z\sqrt{\frac{p(1-p)}{n}}, p + z\sqrt{\frac{p(1-p)}{n}}\right]$$

wherein z is the critical value from the normal distribution.

Problem 1: Foobartendr.io tested four different UX designs for arranging the top 4 cocktails in the following table:

Experiment	Visit	Purchase
#1	24	3
#2	180	30
#3	250	50
#4	100	15

Based on how often customers purchased drinks in each design, can you tell which customer experiment has the lowest purchase rate?

This problem is about making judgment and we need to justify our decision and show how we reach that decision. A first attempt can be to compare simple ratios

Experiment	Visit	Purchase	Ratio
#1	24	3	0.125
#2	180	30	0.167
#3	250	50	0.200
#4	100	15	0.150

It looks like Experiment #1 has the lowest ratio though it has the least data points so maybe it is not valid. If we want to make this judgment with some confidence, then we need to calculate confidence intervals for the proportions. We can construct a 95% confidence interval for an adjusted proportion p in an attempt to account for the lower data points in Experiment #1.

$$\text{adjusted}(p) = \frac{\#\text{ Purchase} + 2}{\#\text{Visit} + 4}$$

and the confidence interval is then

$$p \pm 1.96\sqrt{\frac{p(1-p)}{\#\text{ Visit}}}$$

We get the following confidence interval estimates

Experiment	Visit	Purchase	Adjusted Ratio	Lower bound	Upper bound
#1	24	3	0.178	0.036	0.320
#2	180	30	0.173	0.119	0.228
#3	250	50	0.204	0.155	0.254
#4	100	15	0.163	0.092	0.234

We observe that the calculated confidence intervals overlap and there is no clear loser among the four designs.

This is taken from Paco Nathan slides 'Lies, Damn Lies, Stats'

2.4 Bootstrapping estimation

You may want to escape this section on 'Bootstrapping' for the time being as we will come back to this methods when studying Random forests later on.

To estimate a confidence interval in the case where the population cannot be assumed to normally distributed, we can use the *bootstrapping method*. This method relies on **repeated re-sampling with replacement** of the initial sample as the basis of estimation. An example of bootstrap method to estimate the population mean consists of the following steps:

1. Select a random sample of size n from a population of size N.
2. Re-sample the initial sample by selecting n values with replacement from the initial sample, and compute the sample means for this re-sample.
3. Repeat the step 2 m number of times to produce m re-samples and their associated means.
4. Construct the distribution of the obtained m sample means.
5. Sort the obtained sample means in increasing order.
6. Find the values that exclude the smallest $\alpha/2 \times 100\%$ of means and the largest $\alpha/2 \times 100\%$ of the means. These values become the lower and upper limits of the bootstrap confidence interval estimate of the population mean with $(1\alpha)\%$ confidence.

This gives us an algorithm that can be implemented in a programming language such as Python in order to carry out the bootstrapping method.

Bootstrap Example: Consider the original dataset formed by

						Mean
1	2	3	4	5	6	3.5
5	2	3	1	4	**5**	3.33
4	2	**4**	3	**4**	1	3.00
6	**2**	4	5	3	**2**	3.67
6	**6**	**1**	2	3	**1**	3.17
...						...

We repeat this process say 30 times to get 30 mean values to be sorted and get the confidence interval as described in the algorithm.

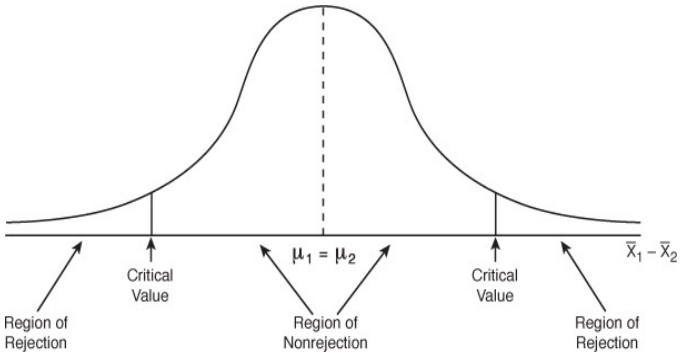
3 Hypothesis Testing

In science, we state a hypothesis about a phenomenon and then we prove it by using investigation and testing. In statistics, *hypothesis testing* is an inferential method designed to evaluate a claim made about a parameter of a population by using a sample statistic.

3.1 The Null and Alternative hypotheses

The *null* hypothesis is the initial premise or the statement a population parameter is equal to a specific value. For example, 'the population mean time to mark exam scripts was 5 days last year'. To evaluate this claim, we will use a sample statistic to estimate the mean and check if it is identical to the population mean. The null hypothesis is usually denoted by H_0. For example we can have $H_0 : \mu = 5$ or $H_0 : \mu_1 = \mu_2$. The *alternative* hypothesis is what is proposed as an alternate premise usually in contradiction to the null hypothesis. For example, $H_1 : \mu \neq 5$ or $H_1 : \mu_1 \neq \mu_2$.

We rely upon the Central Limit Theorem to obtain a normal or t-distribution and the significance level α to divide the distribution into two regions, a region of *rejection the critical region* and a region of *non-rejection*. Note that this coincides with the calculation of confidence intervals.



We distinguish two types of decision-making errors:

- *Type I Error*: This is the error that occurs if the null hypothesis H_0 is rejected when it is true and should not be rejected. The probability of committing a type I error is controlled by the specified significance level α whose most common values are $0.01, 0.05, 0.10$ as stated earlier.
- *Type II Error*: This is the error that occurs if the null hypothesis H_0 is not rejected when it is false and should be rejected. The probability for this type of error is controlled by a parameter β, which depends on the difference between the population and the sample parameters. A small difference leads to a large probability β. The value $1 - \beta$ is known as the *power of the test*.

Risk Trade-off: The two types of errors have an inverse relationship. When you decrease α, you always increase β and vice versa. One way of decrease β without affecting α is to increase the sample size.

3.2 Performing the test: z-test or t-test?

When you perform a hypothesis test, you should follow the steps of hypothesis testing in this order:

1. State the null hypothesis, H_0, and the alternative hypothesis, H_1.

2. Evaluate the risks of making type I and II errors, and then choose the level of significance, α, and the sample size.

3. Determine the appropriate test statistic (t-test or z-test) and the sampling distribution to use as carried out in calculating confidence intervals and identify the critical values that divide the rejection and non-rejection regions. Recall that the z-test uses a normal distribution and the t-test a t-distribution.

4. Collect the data, calculate the appropriate test statistic, and determine whether the test statistic falls into the rejection or the non-rejection region.

5. Make the proper statistical inference. Reject H_0 if the test statistic falls into the rejection area. Otherwise, dd not reject H_0.

3.3 The p-value approach

To carry out a hypothesis testing, we can use the p-value approach in lieu of confidence intervals. The p-value represents the probability of computing a test statistic greater or equal to the sample results given that the null hypothesis H_0 is true. It represents the smallest level at which the null Hypothesis can be rejected for a given set of data. By using the p-value approach, the decision rules are as follows:

- If the p-value is greater or equal to the significance level α, do not reject H_0.
- If the p-value is less than α, reject H_0.

Problem 2: Let us see an example of hypothesis testing now. A famous pizza place claims that their mean delivery time is 20 minutes with a standard deviation of 3 minutes. An independent market researcher claims that they are deflating the numbers for market gains and the mean delivery time is actually more. For this, he selected a random sample of 64 deliveries over a week and found that the mean is 21.2 minutes. Is his claim justified or the pizza place is correct in their claim?

Assume a significance level of 5%. First things first, let us define a null and alternate hypothesis:
$H_0 : \mu = 20$ (What the pizza guy claims)
$H_1 : \mu > 20$ (What researcher claims)
$\sigma = 3, n = 64, \mu_1 = 21.2, \alpha = 0.05$
Let us calculate the Z-value:

$$Z = (21.220)/(3/\sqrt{64}) = 3.2$$

Looking at the standard normal table for this Z-value, we find out that this value has an area of 0.9993 to the left of it; hence, the area to the right is $1 - .99931$, which is less than 0.05. Hence, p-value $< \alpha$. Thus, the null hypothesis is rejected.

Due to the symmetry and nature of the normal distribution, hypothesis tests can be right-tailed like in this case, left-tailed, or two-tailed.