Class Example 1.

The Central Limit Theorem and the Sampling Distribution of the Sample Mean

A captain of a city cat ferry is a hobbyist statistician. Every time he passes next to 'Brisbane Eye' in South Bank, he records the clockwise angular position of a specific unique damaged carriage. He also liked trigonometry in high school, so his recordings are in radians.

Denote his recordings after N passes as, $X_1, \ldots, X_N$ (note each observation recorded is a capital - as each observation is a random variable).

(a) Argue why it is perhaps sensible that $X_1, \ldots, X_N$, are i.i.d. (independent and identically distributed) and uniformly distributed on $[-\pi, \pi]$. What would be reasons for violating this assumption?

(b) What is $E[X_i]$ and $Var(X_i)$?

(c) Draw the pdf of X_i.

```
using PyPlot
x = [-pi,-pi,pi,pi]
y = [0.,1/(2*pi),1/(2*pi),0]
ylim(0.0,0.2)
xlabel("Angle (radians)")
ylabel("PDF (C)")
PyPlot.plot(x,y,label="Eye_PDF");
```

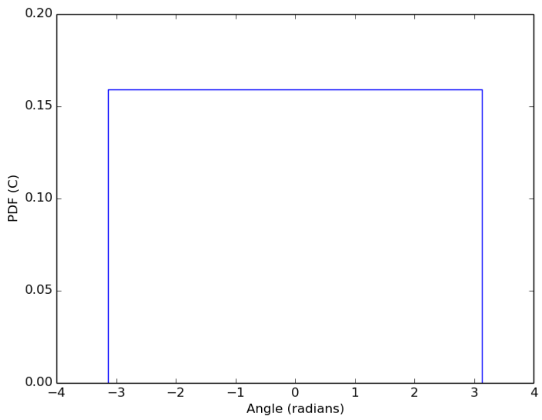

Figure 1: Probability density function of the Brisbane Eye

(d) Assume that the captain passes next to the eye N times a day (each day has a different N), for $N = 2, 3, 4, 5, 10, 20$ and then calculates the sample means:

$$\overline{X}_N = \frac{1}{N} \sum_{i=1}^{N} X_i.$$

```
using Distributions
N=[2,3,4,5,10,20];
[mean(rand(Uniform(-pi,pi),N[_])) for _ in 1:length(N)]
```

(e) Calculate $E[\overline{X}_N]$ and $Var(\overline{X}_N)$ for each N.

(f) Use Monte-Carlo to draw estimates of the pdfs of X_N for each N, each time with 10,000 samples.

```
using Distributions
base = 320
N = [2,3,4,5,10,20]
for i in 1:length(N)
  subplot(base+i)
  PyPlot.plt[:hist]([mean(rand(Uniform(-pi,pi),N[i]))
    for _ in 1:10000],100)
end
```

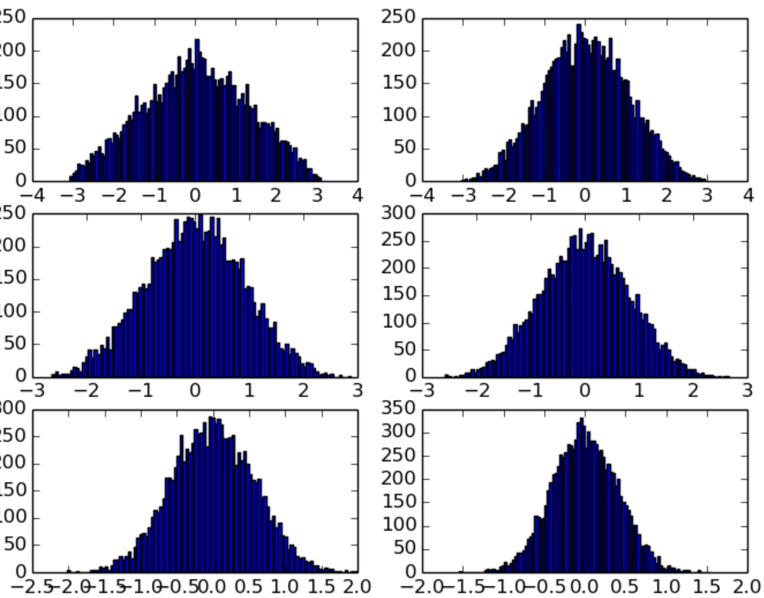

Figure 2: Probability density function of the Brisbane Eye

(g) Comment on how this illustrates the Central Limit Theorem.

(h) Assuming that the uniform distribution assumption is wrong, but the mean and variance estimates in (b) are the same. Would the distribution of $\overline{X}_N$ for large N (e.g. 20) change?

Class Example 2. Confidence intervals and Hypothesis test for large samples

Consider the data file, class4_2.csv containing 100 observations.

(a) Load the data file and carry out a descriptive statistic summary. What is the sample mean? What is the sample standard deviation?

(b) Determine (large sample) confidence intervals of confidence levels: 90%, 95% and 99%.

(c) You now wish to test the Hypothesis: $H_0 : \mu = 203$ vs. $H_1 : \mu \neq 203$. Carry out a hypothesis test (z-test), and find the associated p-value. What are your conclusions with $\alpha = 0.1, 0.05, 0.01, 0.005$?

Class Example 3. Confidence intervals and Hypothesis test for small samples

Consider the data file, class4_3.csv containing 10 observations.

(a) Load the data file and carry out a descriptive statistic summary. What is the sample mean? What is the sample standard deviation?

(b) Determine confidence intervals of confidence levels: 90%, 95% and 99%.

(c) You now wish to test the Hypothesis: $H_0 : \mu = 30$ vs. $H_1 : \mu < 30$. Carry out a hypothesis test (t-test). What are your conclusions with $\alpha = 0.1, 0.05, 0.01, 0.005$?

Question 1. Seeing the CLT with Simulation

Consider the following random variables:

$$\begin{aligned} U &\sim \text{Uniform}(5, 15) \\ V &\sim \text{Exponential}(10) \\ W &\sim \text{Binomial}(10, 0.4) \end{aligned}$$

(a) What is the mean and variance of each?

(b) Consider now,

$$S_n = \sum_{i=1}^{n} X_i,$$

where X_i is either U_i, V_i or W_i (distributed as U, V or W) and different X_i are assumed independent. What is the mean and variance of this random sum, S_n, (a function of n)? Answer this separately for U, V and W.

(c) For X_i either U_i, V_i or W_i, define,

$$\tilde{Z}_n = \frac{S_n - E(S_n)}{\sqrt{var(S_n)}}.$$

Use the CLT to postulate the distribution of $\tilde{Z}_n$ for non-small n.

(d) Generate Monte Carlo estimates of $P(\left|\tilde{Z}_n\right| > 2.0)$ using no less than 10^6 generations of $\tilde{Z}_n$ for every n, (separately for each U, V or W). Compare your results to $P(|Z| > 2.0)$ taken from a normal distribution table, where Z is a standard normal random variable. Do this for $n = 2, 5, 10, 15, 20$. Tabulate your results neatly and explain your results.

Question 2. Sample Mean and Sample Variance

Suppose that a sample of size $n = 20$ is selected at random from a normal population with mean 100 and standard deviation 8. Let $\bar{X}$ be the sample mean and S^2 the sample standard deviation.

(a) Calculate $P(98 \leq \bar{X} \leq 102)$.

(b) Find x such that $P(\left|\bar{X} - 100\right| > x) = 0.01$.

(c) Use Monte-Carlo simulation with 10^5 samples to verify that $E[S^2] = 64$. Then estimate: (i) $P(|S^2 - 64|) > 2$. (ii) $\text{var}(S^2)$.

Question 3. Choice of Sample Size

A normal population has a mean 30 and standard deviation 5. How large must the random sample be if you want the standard error of the sample average to be less than 0.5? Verify your result using Monte-Carlo simulation.

Question 4. Polymer Elasticity

The elasticity of a polymer is affected by the concentration of a reactant. When low concentration is used, the true mean elasticity is 65, and when high concentration is used, the mean elasticity is 75. The standard deviation of elasticity is 6 regardless of concentration. If two random samples of size 25 are taken, find the probability that $\overline{X}_{high} - \overline{X}_{low} \geqslant 2$.

Question 5. Building up Confidence

For a normal population with known variance σ^2, answer the following questions:

(a) What is the confidence level for the interval
$\bar{x} - 2.1\sigma/\sqrt{n} \leqslant \mu \leqslant \bar{x} + 2.1\sigma/\sqrt{n}$?

(b) What is the confidence level for the interval
$\bar{x} - 2.39\sigma/\sqrt{n} \leqslant \mu \leqslant \bar{x} + 2.39\sigma/\sqrt{n}$?

(c) What is the confidence level for the interval
$\bar{x} - 1.85\sigma/\sqrt{n} \leqslant \mu \leqslant \bar{x} + 1.85\sigma/\sqrt{n}$?

(d) What is the confidence level for the interval
$\bar{x} - 1.96\sigma/\sqrt{n} \leqslant \mu \leqslant \bar{x} + 1.96\sigma/\sqrt{n}$?

Question 6. Beverage Machine

A postmix beverage machine is adjusted to release a certain amount of syrup into a chamber where it is mixed with carbonated water. A random sample of 20 beverages was found to have a mean syrup content of $\bar{x} = 0.9$ fluid ounce and a standard deviation of $s = 0.011$ fluid ounce. Find a 99% CI on the mean volume of syrup dispensed. State any assumptions that are made.

Question 7. P-Value

For the hypothesis test $H_0 : \mu = 10$ against $H_1 : \mu > 10$ and variance known, calculate the P-value for each of the following test statistics:

(a) $z = 2.05$

(b) $z = -1.84$

(c) $z = 0.4$

(d) $z = 0$

(e) $z = -2.05$

(f) $z = 3$

Now repeat (a)–(f) when the alternative hypothesis is $H_1 : \mu \neq 10$.

Question 8. Sodium Content in Organic Cornflakes

The sodium content of twenty 300-gram boxes of organic cornflakes was determined. The data (in milligrams) is contained in (*9-65.csv*).

(a) Can you support a claim that mean sodium content of this brand of cornflakes differs from 120 milligrams? use $\alpha = 0.05$, state your hypothesis clearly, find the P-value and make a conclusion.

(b) Check that sodium content is normally distributed (e.g. using the code for Normal probability plots from Assignment 3).

(c) Compute the power of the test if the true mean sodium content is 130.5 milligrams.

(d) What sample size would be required to detect a true mean sodium content of 130.1 milligrams if you wanted the power of the test to be at least 0.75? Explain your answer.

(e) Explain how the question in part (a) could be answered by constructing a two-sided confidence interval on the mean sodium content.