Day 2 Basic statistics

2.1 Statistics

2.1.1 Functions of random numbers

- Consider repeated samples from the same distribution X_1, $X_2, \cdots, X_n$.
- A statistic is some function of these measurements. A commonly used statistic is the sample mean

$\overline{X} = \frac{1}{n} \sum_{k=1}^{n} X_n$

- This is a random number with some distribution, for repeated draws.
- What are the convergence properties of this quantity when there are many samples taken?

2.1.2 Law of large numbers

- The sample mean is a $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_n$.
- Law of large numbers states that

$$\overline{X}_n \to \mu$$

- The characteristic function of a random variable is $\phi_X(t) = \langle e^{itX} \rangle$
- Reminder: for the sum $X = X_1 + X_2$ the characteristi functions are $\phi_X(t) = \phi_{X_1}(t)\phi_{X_2}(t)$
- As these are indpendent samples we have

$$\phi_{\overline{X}}(t) = \langle e^{itX/n} \rangle^n = [\phi_X(t/n)]^n = [1 + \tfrac{it}{n}\langle X \rangle + O(1/n^2)]^n = e^{it\langle X \rangle + O(}$$

- But $e^{it\mu}$ is the characteristic function of a sharply-peaked distribution (Dirac Delta) around μ.

2.13 Central limit theorem and standard error on the mean

- Go beyond the first order in the t expansion of the characteristic function of $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_n$

$$\phi_{\overline{X}}(t) = \langle e^{itX/n} \rangle^n = [\phi_X(t/n)]^n = [1 + \tfrac{it}{n}\langle X \rangle - \tfrac{t^2}{2n^2}\langle X^2 \rangle + \cdots]^n \simeq e^{[}$$

- **Question:** Calculate the argument of the exponential to order $1/n$.
  Compare it to the list of characteristic functions is 1.4.3. Infer the distribution and give its mean and variance.

2.1.4 Question: Examples of the central limit theorem

- Consider two gamma distributions (α, β) with shape factors $(0.5, 0.5)$ and $(5.0, 5.0)$.
- What are their means?
- Plot the pdfs for these two cases.
- Consider a sample size n for calculating the sample mean. Plot the histogram of sample means for these two cases for N repeated experiments, one graph for each gamma case.
- Plot the normal distributions corresponding to the central limit theorem on the same graphs.

2.2 Sample statistics

2.2.1 Sample mean

- The sample mean tends towards the true mean as $n \to \infty$. Is it a biased estimator?

$\overline{X} = \frac{1}{n} \sum_{k=1}^{n} X_n$

- What is the typical value of the sample mean? What is it's value on average?

$\langle \overline{X} \rangle = \frac{1}{n} \sum_{k=1}^{n} \langle X_n \rangle$

- This is just μ so it is **not** a biased estimator.
- What about the sample variance?

2.2.2 Sample variance

- The sample mean is $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_n$ and $\langle \overline{X} \rangle = \mu$.
- The variance of the population is Var$(X) = \langle X^2 \rangle - \langle X \rangle^2$ and call this σ^2
- The sample variance is defined as $s_b^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^2$ (Note: use of subscript b explained later)
- We need to use the sample variance to infer the population variance so we can estimate the standard-error on the mean.
- What is the sample variance expectation? Is $\langle s_b^2 \rangle = \sigma^2$ or not?

2.2.3 Question: Sample variance derivation

- Using the sample mean and variance

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_n \quad \text{and} \quad s_b^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^2$$

- Express the following variance in terms of the population $\left\langle X^2 \right\rangle - \langle X \rangle^2 = \sigma^2$

$$\left\langle s_b^2 \right\rangle = \frac{1}{n}\sum_{k=1}^{n} \left\langle \left(X_k - \frac{1}{n}\sum_{j=1}^{n} X_j \right)^2 \right\rangle$$

2.2.4 Question: Numerical test

- Generate some random numbers
- Calculate the sample mean
- Calculate the biased sample variance and compare to the results using the **var** command.
- Does **Julia** use the a biased or unbiased variance?

2.3 Distribution of sample variance

2.3.1 Distribution of a squared normal

- A random number Z drawn from a standard normal (zero mean, unit variance).
- We are interested in the statistics of its square

$$X = Z^2$$

- **Question:** Using the transformation rules, derive the distribution for X.

2.3.2 Sums of squared normals

- We now consider a sum of k squared normals

$$Q_k = \sum_{j=1}^{k} Z_k^2$$

- What does this distribution look like?
- We first compare the χ_1^2 distribution for one Z^2 with the gamma distribution form

$$\chi_1^2(x) = \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2} \quad \text{and} \quad f(x) = \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} e^{-\beta x}$$

- The χ_1^2 distribution looks like a gamma with $\alpha = 1/2$ and $\beta = 1/2$.
- Summation rule for gamma distributions is that the αs add if the βs are the same.
- The distribution for Q_k is therefore

$$\chi_k^2(x) = \frac{1}{\Gamma(k/2)} \frac{1}{2^{k/2}} x^{k/2-1} e^{-x/2}$$

- This is a χ_k^2 distribution with k degrees of freedom.

2.3.3 Sample variance distributions for normals

- The sample mean and unbiased variance

$$\overline{X} = \frac{1}{n} \sum_{k=1}^{n} X_n \quad \text{and} \quad s_u^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \overline{X})^2$$

- The sample mean has distribution for large n that approaches normal with μ, σ^2/n.
- What is the distribution of the sample variance?
- Can be calculated for Normally distributed random variables.
- This calculation is beyond this course's scope (see Cochran's theorem).
- Result is that, the ubiased sample variance s_u^2 and population variance σ^2 obey

$s_u^2(n-1)/\sigma^2$ follows a χ_{n-1}^2 distribution.

2.3.4 Question: sample variance numerics

- Draw n samples and calculate the sample mean and unbiased sample variance for normally distributed random numbers. Repeat this a large number of times and plot the histograms of the sample means and sample variances, together with their theoretical predictions.

2.4 Summary and additional questions

Day 2 Basic statistics

Additional questions

Q2.4.1 Sample mean and variance for house prices
Q2.4.2 Distribution of house prices

Q2.4.1 Sample mean and variance of house prices

- Go to the site https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads (https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads)
- Download the current month (September 2018 data) in csv format onto your computer.
- Take a sample of n house prices and calculate the sample mean and variance.
- Plot what you expect the distribution of the mean to be using these results.
- Show that the sample mean has a Normal distribution (for sufficiently large n).
- Calculate the population mean and variance for house prices less than or equal to £2M.

Q2.4.2 Distribution of house prices

- Plot a histogram of the house prices that are less than or equal to £2M.
- Analyse the data. Which, if any, distribution provides a good fit of the data?
- HINT: find the command that gives you a histogram in a vector format.

In [ ]:

```

```