Probabilities & Data

Week 8: Continuous Distributions II

DR NICK FEWSTER-YOUNG

Topics ☺

- Calculating some probabilities from last week ☺
- Simulations in R
- We will look at discretising a continuous random variables so that we can simulate a continuous random variable.
- Central Limit Theorem & Confidence Intervals
- Approximation to the Normal Distribution

- Next week – Bayesian Statistics

Example

Let’s kick off with an example to test our knowledge from last week:

$$f_X(x) = \begin{cases} -k, & 0 \leq x < 2, \\ 2k\,, & 2 < \ x \leq 3, \\ 0, & otherwise \end{cases}$$

(a) Find k such that f_X is a probability density function.

(b) Derive the cumulative distribution function.

(c) Use R to calculate $P(1 \leq X \leq 2.5)$.

(d) Use R to compute E(X).

Example

$$f_X(x) = \begin{cases} k, & 0 \leq x < 2, \\ 2k\ , & 2 < \ x \leq 3, \\ 0, & otherwise \end{cases}$$

(a) By drawing a picture of f then we see that $k = \frac{1}{4}$ since the area under the curve is 4k and we must have 4k=1.

(b) See that the cumulative function is then given by

$$F_X(x) = \begin{cases} 0, & x \leq 0, \\ \frac{x}{4}, & 0 \leq x \leq 2 \\ \frac{x}{4} + \frac{x-2}{4}\ , & 2 < \ x \leq 3, \\ 1, & x \geq 3 \end{cases}$$

Example

$$F_X(x) = \begin{cases} 0, & x \le 0, \\ \frac{x}{4}, & 0 \le x \le 2 \\ \frac{x}{4} + \frac{x-2}{4}, & 2 < x \le 3, \\ 1, & x \ge 3 \end{cases}$$

(c) By using the cumulative distribution function, we can compute the (otherwise use R)

$$P(1 \le X \le 2.5) = F(2.5) - F(1) = (0.625 + 0.125) - 0.25 = 0.5.$$

(d) By using R, we have

$$E(X) = \frac{1}{4} \times \frac{4}{2} + \frac{1}{2} \times \frac{9-4}{2} = 1.75.$$

Chebyshev's Inequality

Theorem *(Chebyshev's Inequality)* For any positive constant $a > 0, \text{and } X$ *is a random variable with bounded variance then*

$$P(|X - E(X)| \geq a\ ) \leq \frac{Var(X)}{a^2}$$

Recall the above inequality from the Discrete case for finding upper limits of the probability for certain values.

Example

Suppose X is a continuous random variable which has an Uniform distribution over the region [-1,1]. By definition this implies that

$$f_X(x) = \frac{1}{2} \text{ over } -1 \le x \le 1$$

and 0 elsewhere. This tells us the mean of X, $E(X) = 0$ and $Var(X) = \frac{1}{3}$.

By using Chebyshev Inequality, we have

$$P\left(X \le -\frac{\sqrt{3}}{2}\ \ OR\ X \ge \frac{\sqrt{3}}{2}\right) = P\left(|X| > \frac{\sqrt{3}}{2}\right) \le \frac{\frac{1}{3}}{\frac{3}{4}} = \mathbf{\frac{4}{9}}$$

We could use R to get a more accurate answer by

- 1-punif(sqrt(3)/2,-1,1) + punif(-sqrt(3)/2,-1,1)

[1] 0.1339746

Example

The time taken in hours for the next 222 bus to arrive at a Mawson Lakes bus-stop is distributed as exponential with $\lambda = 5$, that is, the average waiting time after having just missed a 222 bus and having to wait for the next one at the bus-stop is $E[X] = 1/5 = 0.2$ hours.

Question: What is the probability of having to wait more than the average waiting time?

Answer: This is $P(X > 0.2)$, thus by using R, we have

$P(X > 0.2) =$

```
> 1 - dexp(0.2, rate = 5)

[1] 0.3678794412
```

Example (cont.)

Now what happens if we want to track the time to the 2nd arrival?

We assume that immediately after the 1st arrival of the event, the distribution for the arrival to the 2nd time the event occurs is identical and independent to that of the 1st event.

So now we have both X_1 and X_2 are Exp(λ) and we want to find the distribution of $S_2 \coloneqq X_1 + X_2$.

To find S_2, we need to use the formal definition for finding the distribution of S_2:

Theorem. If X_1, X_2 are two random variables with probability distributions f_{X_1}, f_{X_2} respectively then

$$f_{S_2}(s) \coloneqq \int_{-\infty}^{s} f_{X_1}(x) \times f_{X_2}(s-x)\, dx$$

Example (cont.)

For the example before, the probability density function for the sum is given by

$$f_{S_2}(s) = \lambda^2 s\, e^{-\lambda s}$$

and if $s < 0$ then the distribution is 0.

Calculating Probabilities of the Exponential

In applications when the exponential distribution is used then we can actually use the Gamma distribution in **R** since the exponential distribution with parameter λ is equal to the Gamma(1,λ).

Also, if we were to have $X_1, X_2, \dots X_n$ random variables with exponential distributions then

$$S_n \coloneqq X_1 + X_2 + \cdots + X_n$$

Can be modelled with Gamma(1, n, λ).

The **R** commands are "dgamma(x,n,λ)", "pgamma(x,n,λ)" and "rgamma(N,n,λ)" for calculating pdf, cdf and generating random numbers respectively.

Example

Example. Suppose the times taken in hours for each 222 bus, immediately after the previous one, to arrive at a Mawson Lakes bus-stop are independently and identically distributed as exponential with $\lambda = 5$.

What is the distribution of the time taken until the 4th 222 bus arrives?

Let S4 denote the time taken for the 4th 222 bus. That is $S_4 \sim Gamma(4,5)$.

What is the probability that the 4th 222 bus arrives after 1 hour?

Solution: Thus,

$P(S_4 > 1) =$

> $1 - pgamma(1,4,5)$

[1] 0.2650259

Example

Example. Using the same information for the 222 buses as in the example above in 4.16. Simulate the arrival times of the buses within a 2 hour interval.
Solution. We need to ensure that we have enough arrivals to cover a 2 hour interval, so perhaps try 10 first. We may need more if the arrival time of the 10th bus still falls short within the 2 hour interval. Note that each X_i is distributed as Exp(5)≡Gamma(1,5).

```
> X <- rgamma(10,1,5)
```

```
[1] 0.268815449 0.309877360 0.007159928 0.241296355 0.008835665 0.184050547
0.283352856 0.074770638 0.358235268 0.470780165
```

```
> cumsum(X) [1] 0.2688154 0.5786928 0.5858527 0.8271491 0.8359848 1.0200353 1.3033882
1.3781588 1.7363941 2.2071742
```

X displays the individual arrival times. cumsum(X) adds them up con- secutively so that you can check whether you have enough arrivals beyond 2 hours. So in my case, my 9th arrival is at 1.73 hours and my 10th arrival arrives after 2 hours at 2.21 hours, so I have 9 arrivals within 2 hours.

Normal Probabilities in R

In **R** , the commands are “dnorm(x,μ,σ)”, “pnorm(x,μ,σ)” and “rnorm(N,μ,σ)” for density, calculating cumulative probability and for generating random numbers respectively.

Example. The distribution of the heights (in metres) of a student pop- ulation is normal with mean height 1.75m and standard deviation 0.1m. What is the probability of finding a student whose height is less than 1.65m?

We could use the standard normal table attached or online:

$$\Phi(-1) = P(Z < -1) = 0.1587.$$

Using **R** , we have
> pnorm(1.65,1.75,0.1)

[1] 0.1586553

Simulate some Normal data

Example. Using the same parameter values as in the previous example, simulate the heights of 10 students.

```
> rnorm(10,1.75,0.1)
```

```
[1] 1.802416 1.866994 1.880959 1.608311 1.824251 1.663301 1.757672 1.783826 1.846873
1.760915
```

- The normally distributed random variable also has the property that sums of normally distributed random variables are also normal.
- Let $S_2 \coloneqq Z_1 + Z_2$ where the Z_i are each distributed as N (0, 1) and independent of each other. Then S2 ~ N(0,$\sqrt{2}$)

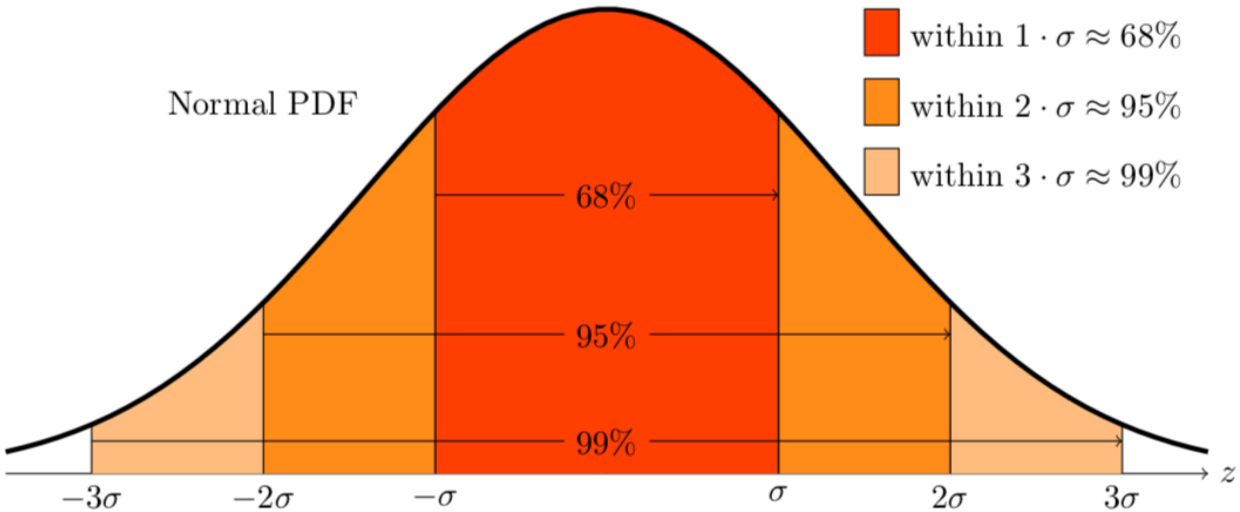
The Normal Table

Normal Curve Areas
Standard normal probability in right-hand tail (for negative values of z, areas are found by symmetry)

	Second decimal place of z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.00135									
3.5	.000 233									
4.0	.000 031 7									
4.5	.000 003 40									
5.0	.000 000 287									

From R. E. Walpole, *Introduction to Statistics* (New York: Macmillan, 1968).

Estimating Normal Probabilities



Law of Large Numbers

- An average of many measurements is more accurate than a single measurement.
- If we are interested in the average of a data sample, then investigating averages of averages is more accurate.
- Let $X_1, X_2, \ldots, X_n$ be identically independent random variables all with the same population mean, μ and standard deviation σ. Let

$$X_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n} \sum X_i.$$

Then for any (small number) a, we have

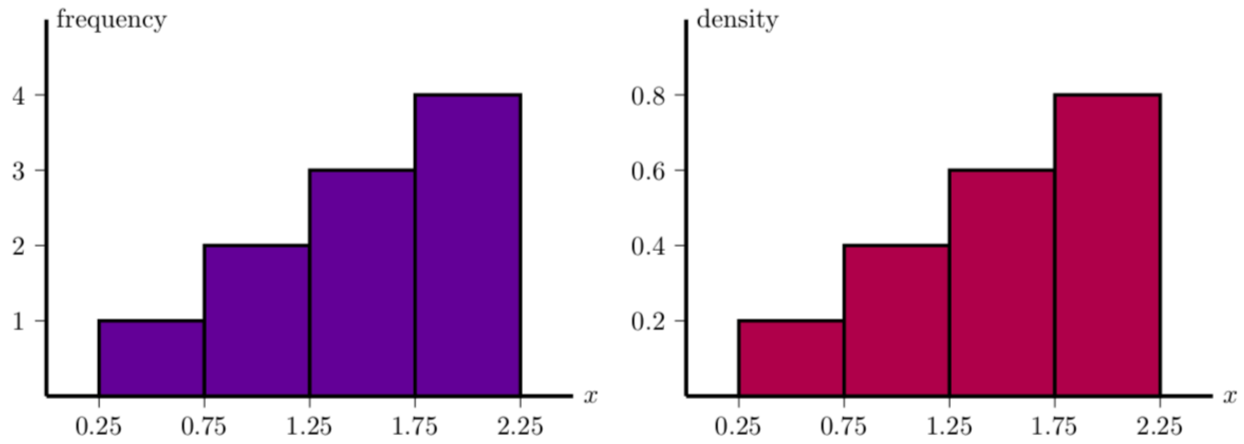
$$\lim_{n \to \infty} P(|\bar{\bar{\bar{X}}}_n - \mu| < a) = 1$$

This says that for large n, the mean of the sample is close to the population mean.

Recall Histograms

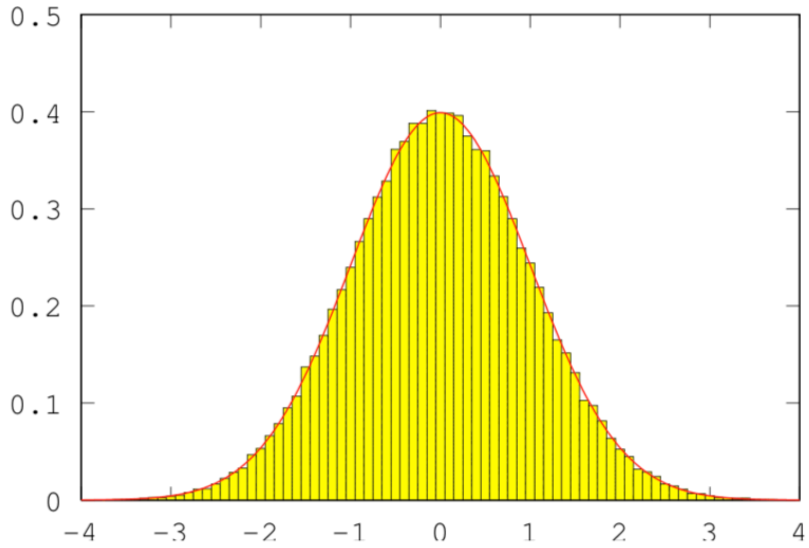
Frequency: gives you the height of a bar over a bin = number of data points in the bin

Density: Area of the bar and relates to the probability density function. So the total area is 1.



Law of Large Numbers & Histograms

By using histograms and using the theory of law of large numbers then the probability density histogram converges to the probability density function.



Central Limit Theorem

- Let $X_1, X_2, \ldots, X_n$ be independent random variables with population mean μ and standard deviation σ. For each n:

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

and

$$S_n \coloneqq X_1 + X_2 + \ldots + X_n$$

Then for large n, we have

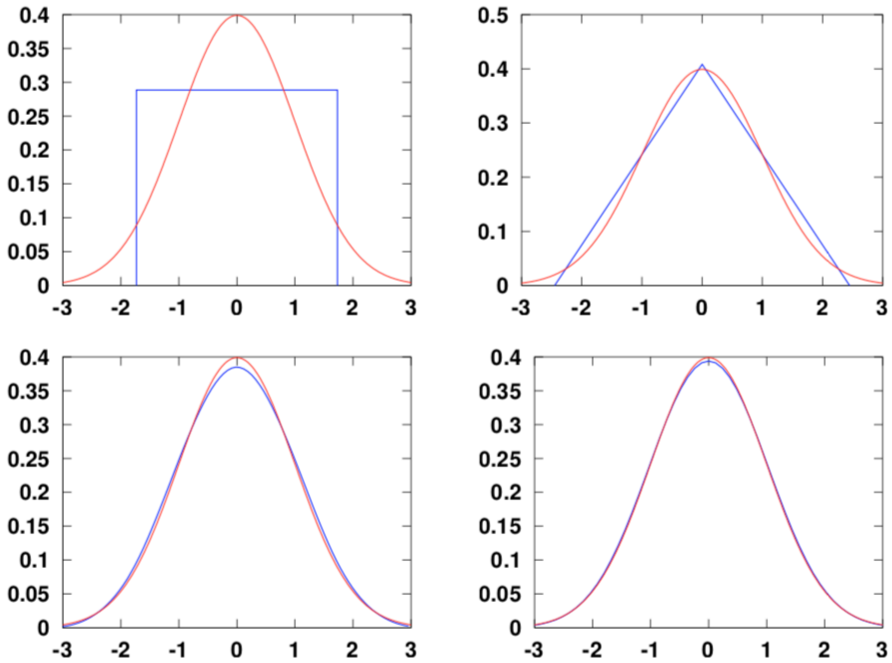
$$\bar{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$S_n \approx N(n\mu, n\sigma^2)$$

$$\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$$

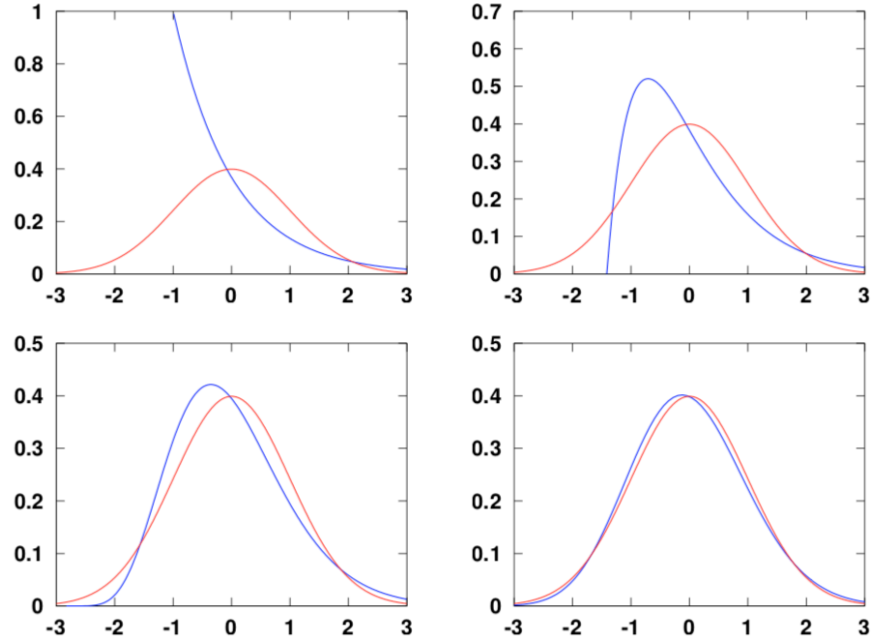
CLT Examples

Some examples of standardised Uniform Distribution for averages of size n = 1,2,4,12.



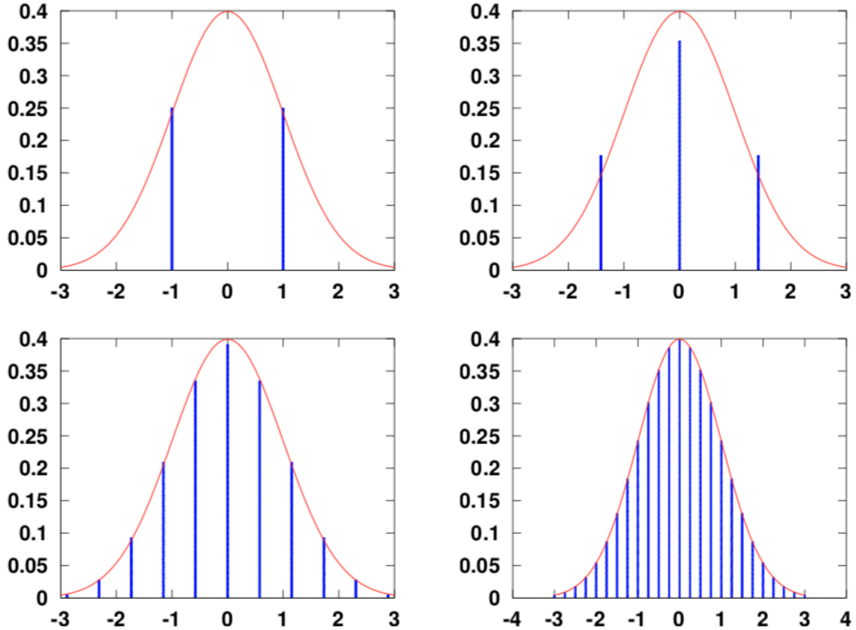
CLT: some examples

Some examples of standardised Exponential Distribution for averages of size n = 1,2,8,64.



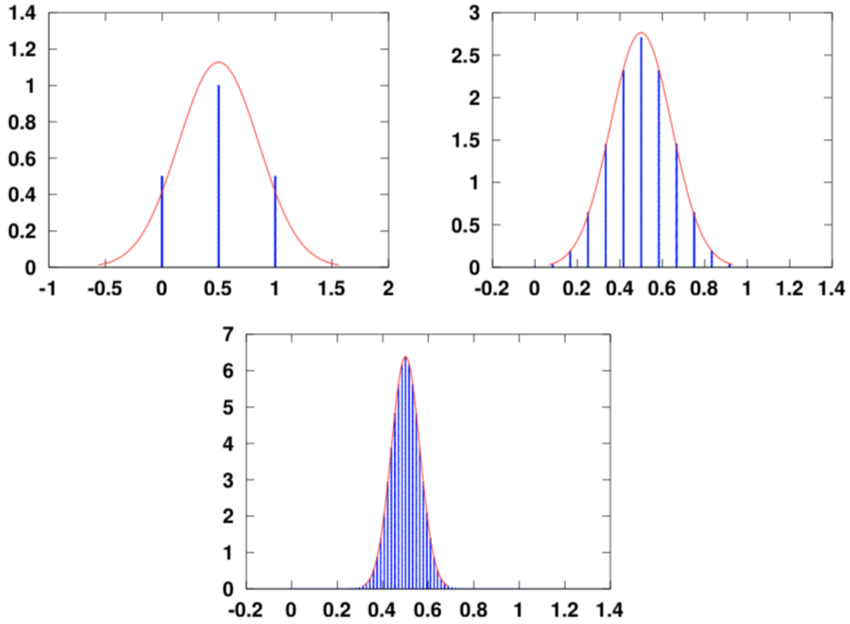
CLT : some examples

Some examples of standardised Bernoulli Distribution with p=0.5 for averages of size n = 1,2,12,64.



CLT : some examples

The (non. Std) average of n Bernoulli with p=0.5 random variables with n=4, 12, 64.



Sampling from the Normal

Let's produce in **R** a single random sample from an approximate standard Normal distribution:

Suppose we roll a 10 sided dice nine times.

R output:

Note that $\mu =, \sigma =$ for this dice.

Now the average of nine rolls is a sample from the average of 9 independent random variables. The CLT says this average is approximately Normal with $\mu = 5.5, \sigma = 2.75$.

Furthermore, if we decided to standardise the average of the 9 rolls then we get

$$z = \frac{\bar{x} - 5.5}{2.75} \approx N(0,1)$$

Generating some Samples

1. Generate a frequency histogram of 1000 samples from an exponential(1) random variable.

2. Generate a density histogram for the average of 2 independent exponential(1) random variable.

3. Using rexp(), matrix() and colMeans() generate a density histogram for the average of 50 independent exp(1) random variables. Make 10000 sample averages and use a binwidth of .1 for this.

Look at the spread of the histogram.

4. Superimpose a graph of the pdf of N(1, 1/50) on your plot in problem 3. (Remember the second parameter in N is σ2.)

Example

I've noticed that taxis drive past City West Campus on the average of once every 10 minutes.

Suppose time spent waiting for a taxi is modeled by an exponential random variable

$$X \sim Exponential\left(\frac{1}{10}\right); \qquad f(x) = \frac{1}{10}e^{-\frac{x}{10}}$$

(a) Sketch the pdf of this distribution.

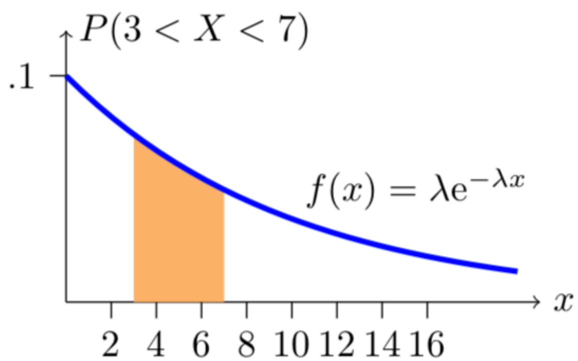
(b) Shade the region which represents the probability of waiting between 3 and 7 minutes .

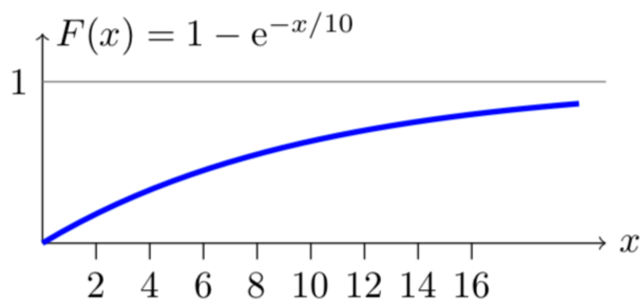
(c) Use compute the probability of waiting between between 3 and 7 minutes for a taxi.

(d) Compute and sketch the cdf.

Solution:

For parts (a), (b), (d) :





(c) By using R, we can calculate the P(3<X<7)

> dexp(7) – dexp(3) = 0.244

Correlation

Recall we looked at correlation in the discrete case in Week 3, we return with the definition for the continuous case.

- The correlation coefficient between two random variables X, Y is defined as

$$\rho = Cor(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

where Cov(X,Y) is the covariance of X and Y.

Some properties are:

- It's a ratio and thus dimensionless
- Its only takes values between $-1 \leq \rho \leq 1$. The closer to 1 or -1 means the stronger the linear relationship (1-1) is between X, Y (or inverse relationship).

Anecdotal Examples

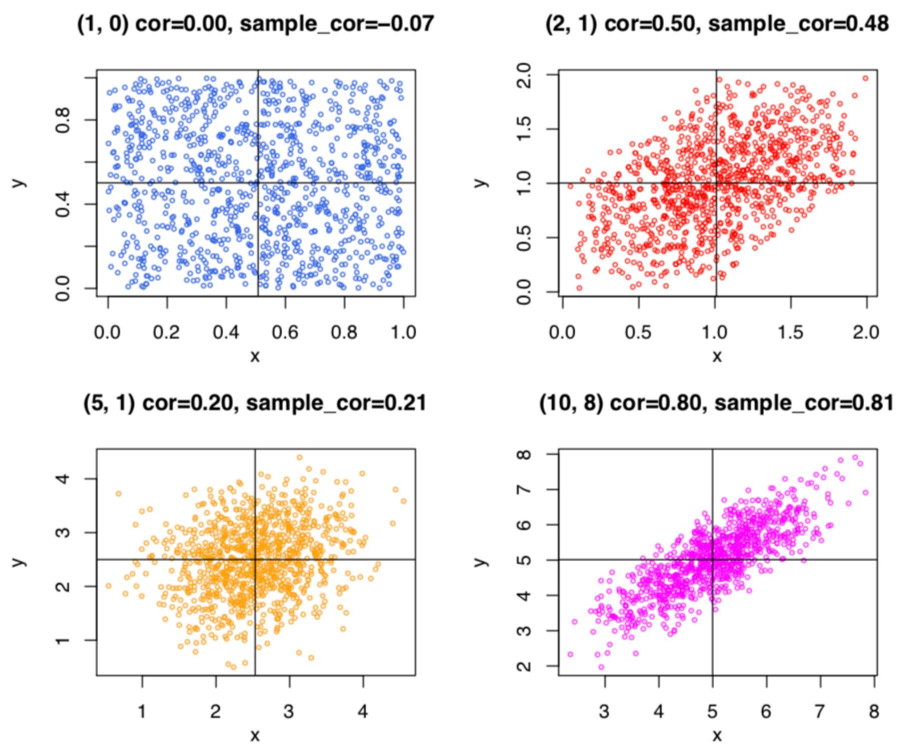
- Over time, amount of Ice cream consumption is correlated with number of pool drownings.
- In 1685 (and today) being a student is the most dangerous profession.
- In 90% of bar fights ending in a death the person who started the fight died.
- Hormone replacement therapy (HRT) is correlated with a lower rate of coronary heart disease (CHD).

Anecdotal Examples

- Ice cream does not cause drownings. Both are correlated with summer weather.
- In a study in 1685 of the ages and professions of deceased men, it was found that the profession with the lowest average age of death was “student.” But, being a student does not cause you to die at an early age. Being a student means you are young. This is what makes the average of those that die so low.
- A study of fights in bars in which someone was killed found that, in 90% of the cases, the person who started the fight was the one who died.

Of course, it’s the person who survived telling the story. :)

Scatter Plots



Approximation of Distributions

- We have looked at approximation theory of distributions to Binomial, Poisson, Hypergeometric and plenty of Discrete Distributions.
- The Central Limit Theorem and Law of Large Numbers allows us to approximate parameters and Discrete Distributions such as the Binomial and Poisson to the Normal Distribution.
- Just remember, the rule of thumb is the shape of the discrete probability density function needs to be bell-shaped (NORMAL).
- Testing if your sample is Normally Distributed….
- P-plots and Q-plots will be your friend here and if the p-value (probability that its not Normal) is above 0.05 then we can assume its population is Normal. MORE on this in weeks to come.

Approximations to Normal Distribution

Binomial Approximation

The normal distribution can be used as an approximation to the binomial distribution, under certain circumstances, namely:

If $X \sim B(n, p)$ and if n is large and/or p is close to ½, then X is approximately $N(np, npq)$

(where $q = 1 - p$).

In some cases, working out a problem using the Normal distribution may be easier than using a Binomial.

Approximations to Normal Distribution

Poisson Approximation

The normal distribution can also be used to approximate the Poisson distribution for large values of λ (the mean of the Poisson distribution).

If $X \sim Possion(\lambda)$ then for large values of λ, $X \sim N(\lambda, \lambda)$ approximately.

Continuity Correction

The binomial and Poisson distributions are discrete random variables, whereas the normal distribution is continuous. We need to take this into account when we are using the normal distribution to approximate a binomial or Poisson using a **continuity correction**.

Approximate Example

So when working out probabilities, we want to include whole rectangles, which is what continuity correction is all about.

Example: Suppose we toss a fair coin 20 times. What is the probability of getting between 9 and 11 heads?

Let X be the random variable representing the number of heads thrown.

$$X \sim \text{Bin}(20, \tfrac{1}{2})$$

Since p is close to ½ (it equals ½!), we can use the normal approximation to the binomial.

$$X \sim N(20 \times \tfrac{1}{2}, 20 \times \tfrac{1}{2} \times \tfrac{1}{2}) \text{ so } X \sim N(10, 5) .$$

We want to $P(9 \leq X \leq 11)$. Notice that the first rectangle starts at 8.5 and the last rectangle ends at 11.5. Using a continuity correction, therefore, our probability becomes $P(8.5 < X < 11.5)$ in the normal distribution.

Finish off the example ☺ in class.

Next Week

- Bayesian Statistics
- Bivariate Distributions
- Conditional Probability