Elementary Statistics: Math 080

Jordan Hanson

July 25, 2022

Whittier College Department of Physics and Astronomy

Summary

Summary

1. Topics from Chapter 4: 4.1 - 4.4
    - Discrete random variables
    - Expectation values and standard deviations
    - The binomial distribution
    - The geometric distribution
2. Topics from Chapter 6: 6.1 - 6.4
    - 2.1 The normal and standard normal distributions
    - 2.2 Using normal distributions

Discrete Random Variables

Discrete Random Variables

A discrete random variable is a property of data that can be counted with integers.

Examples:

- Times a baby eats per day
- Number of students in a class
- The number of wins a team has in a season
- *The number of calories we ate yesterday* - We may think of this as discrete if we have to round to the nearest calorie

Bin	n	$P(x)$	$x * P(x)$	$(x - \mu)^2 P(x)$
0-10k	13			
10-20k	15			
20-30k	20			
30-40k	11			
40-50k	9			
50-60k	9			
60-70k	6			
70-80k	7			
80-90k	5			
90-100k	3			
100k+	2			
Totals	100			

Table 1: Wage data for 100 Los Angeles County workers.

$P(X)$ is a **Probability distribution function** of a discrete random variable. PDFs are tools for answering questions like:

1. What is the probability that a random individual in LA County earns yearly wages in the top 5 categories of Tab. 1?
2. What is the probability that a random individual in LA County earns yearly wages in the bottom 5 categories of Tab. 1?
3. What is the *expectation value* of Tab. 1?
4. What is the *standard deviation* of Tab. 1?

Discrete Random Variables

Bin	n	$P(x)$	$x * P(x)$	$(x-\mu)^2 P(x)$
0	45			
1	190			
2	410			
3	220			
4	80			
5	55			
Totals	1000			

Table 2: Number of cars owned by 1,000 California citizens.

Consider Tab. 2 above.

1. What is the probability that a random Californian has 2 or fewer cars, according to Tab. 2?
2. Suppose a random Californian owns 4 cars. How many standard deviations above the mean is this, according to Tab. 2?

Bin	n	$P(x)$	$x * P(x)$	$(x-\mu)^2 P(x)$
200-300k	110			
300-400k	130			
400-500k	140			
500-750k	270			
750-1000k	100			
1000k+	250			
Totals	1000			

Table 3: Values of 1,000 residential properties in Los Angeles County.

Consider Tab. 3 and Tab. 1 above.

1. Consider the wage distribution of Tab. 1, and consider the home value distribution of Tab. 3. What is the average home value divided by the average yearly wage? What statistical fact does this reveal?
2. Typically, residents of California devote 30-40 percent of their budget to housing. Take 35 percent as a good estimate, and apply it to the prior calculation. How many years must someone work for the average wage to purchase an average home?
3. For more data and interesting figures, see `https://datausa.io/profile/geo/los-angeles-county-ca`

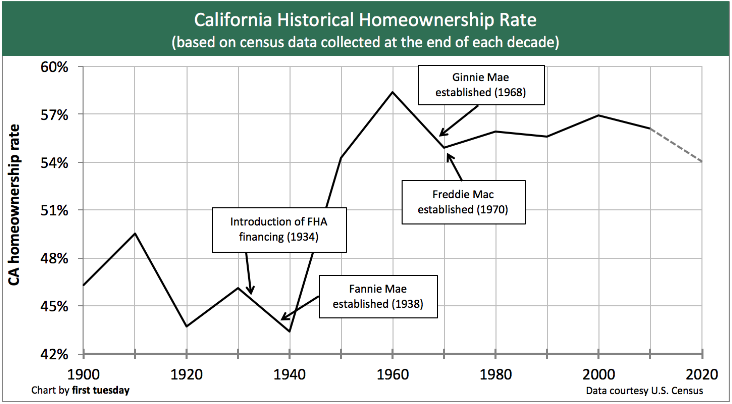

Figure 1: The effect of the FHA on home ownership in California across several decades.

The Binomial Distribution

The Binomial Distribution

Suppose we suspect a discrete random variable data set is binomially-distributed. The mean is $\mu = 2.1$ and the standard deviation is $\sigma = 1.0$. Suppose this data had 10 trials. Independently, someone tells us that the probability of a trial being successful in a similar experiment was 0.4. Is this data binomially-distributed?

- A: Yes, the mean follows $\mu = Np$ within one standard deviation.
- B: No, the mean does not follow $\mu = Np$ within one standard deviation.
- C: Yes, the mean implies a probability of success of 0.4.
- D: No, the standard deviation is too large to make the determination.

Suppose we are working with a biological experiment to predict the behavior of a small aquatic creature when its environment is inside a large magnetic field. The creature can either choose to go left (in the direction of the compass) or right (opposite to the compass). We conduct 10 trials on the same individual, and then repeat on 10 different individuals. How would you determine if the creatures are following the magnetic field?

- A: Count the number of times in all runs, all trials, that the individuals go left. If it is more than half the time, $p > 50\%$.
- B: Count the number of times per run that the individuals go left. Calculate the expectation value of those frequencies, and derive p.

Probability Distributions Can Be Continuous

Suppose instead of *counting trials*, we measure a number. Identify below: discrete random variable or continuous random variable?

1. Measuring the heights of a sample of people
2. Measuring the number of home runs a baseball player earns per season
3. Measuring the number of hands a poker player wins per game won
4. Measuring the wind speed on the top of a mountain

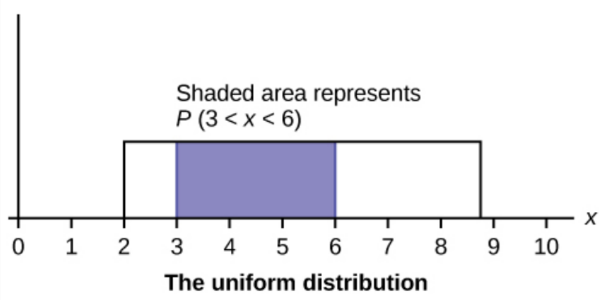

Figure 2: A uniform PDF of a *continuous random variable.*

- How do we *normalize* the frequencies?
- How do we calculate probabilites?
- What are the expectation value and standard deviation?

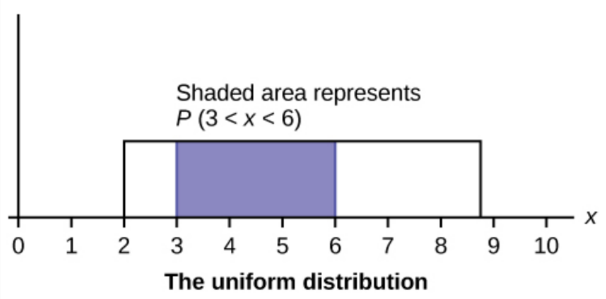

Figure 3: A uniform PDF of a *continuous random variable.*

- Normalization: $1/(b-a)$
- Probability that $x_1 < x < x_2$? $(x_2 - x_1)/(b-a)$
- $E[x] = \mu = (b+a)/2$, $\sqrt{Var[x]} = \sigma = \sqrt{(b-a)^2/12}$

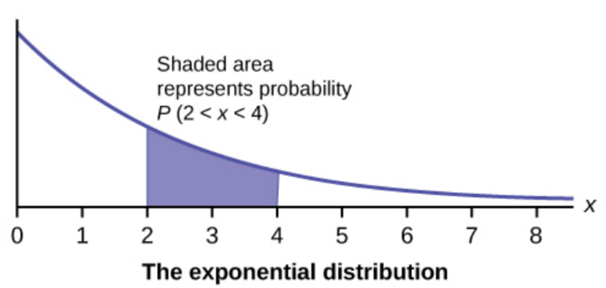

Figure 4: An exponential PDF of a *continuous random variable.*

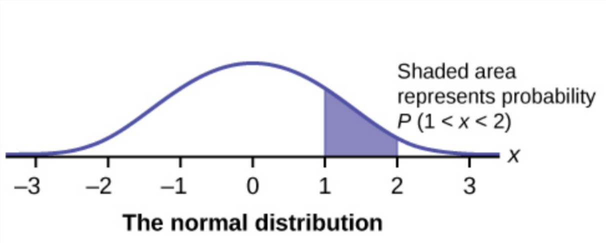

Figure 5: A normal distribution PDF of a *continuous random variable.*

Suppose we record the volume of milk a baby drinks per feeding when between the ages of 0-3 months. The volumes are uniformly distributed between 3.0 and 4.0 ounces per feeding. We record 100 measurements.

- What is the normalization? Or, graph the PDF.
- What is the mean volume?
- What is the standard deviation?
- What is the probability of finding a measurement in the set between 3.2 and 3.5 ounces per feeding?

Suppose we record the volume of milk a baby drinks per feeding when between the ages of 0-3 months. The volumes are uniformly distributed between 3.0 and 4.0 ounces per feeding. We record 100 measurements.

- What are the quartiles of the data?
- What is the 90th percentile of the data?

Interactive Questions

Suppose we are looking at a distribution (histogram/PDF) of a baseball team's batting average (probability a player gets a hit), and it is uniformly distributed. The lowest value is 0.200 and the highest is 0.400. What is the mean of the distribution?

- A: 0.200
- B: 0.300
- C: 0.400
- D: 0.350

Interactive Questions

Suppose we are looking at a distribution (histogram/PDF) of a baseball team's batting average (probability a player gets a hit), and it is uniformly distributed. The lowest value is 0.200 and the highest is 0.400. What is the median?

- A: 0.200
- B: 0.300
- C: 0.400
- D: 0.350

Interactive Questions

Suppose we are looking at a distribution (histogram/PDF) of a baseball team's batting average (probability a player gets a hit), and it is uniformly distributed. The lowest value is 0.200 and the highest is 0.400. What is the probability of being between 0.300 and 0.350?

- A: 10 percent
- B: 15 percent
- C: 25 percent
- D: 50 percent

Statistics and Probability: The Normal Distribution

The *mean*, μ, and *standard deviation*, σ, of a data set $\{x_i\}$ are defined as

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{1}$$

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2 \tag{2}$$

Octave commands:

```
x = randn(100,1);
mean(x)
std(x)
```

One nice theorem: *The variance is the average of the squares minus the square of the average.* Let $\langle x \rangle$ represent the average of the quantity or expression x. We have

$$\sigma_x^2 = \langle x^2 \rangle - \langle x \rangle^2 \tag{3}$$

Proof: observe on board.

Note: There is a distinction between the *process or signal process* and the *the data*. Just because the data has a given μ and σ does not imply that the signal process has or will continue to have the exact same values of μ and σ. The underlying process could be *non-stationary*.

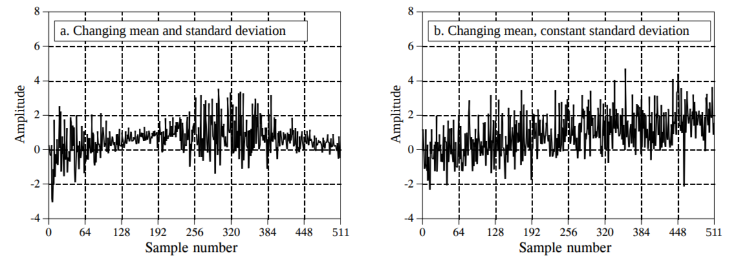

Figure 6: Signal processes in (a) and (b) are considered non-stationary because one or both of μ and σ depend on time.

A histogram is an object that represents the frequency[1] of particular values in a signal. For example, below is a histogram of 256,000 numbers drawn from a probability distribution:

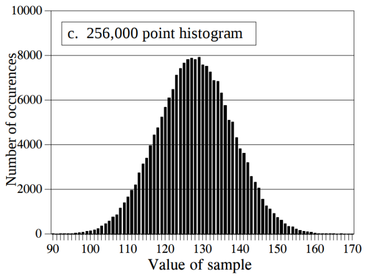

Figure 7: The histogram contains counts versus sample values.

[1]Careful: the word frequency refers to the number of occurences in the data, not a sinusoidal frequency.

The following octave code should reproduce something like Fig. 7 from the textbook:

```
x = randn(256000,1)*10.0+130.0;
[b,a] = hist(x,100);
plot(a,b,'o');
```

The function *randn(N,M)* draws $N \times M$ numbers from a normal distribution and returns them in the size the user desires. The function *hist(x,N)* creates N bins and sorts the data x_i into them.

For data that is appropriately stationary, we can use histograms to estimate μ and σ faster, since we only have to loop over bins rather than every data sample. Let H_i represent the counts in a given bin, and i represent the bin sample. We have:

$$\mu = \frac{1}{N}\sum_{i=1}^{M} iH_i \tag{4}$$

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{M} (i-\mu)^2 H_i \tag{5}$$

To obtain the mean in signal *amplitude*, you'll have to convert bin number to amplitude.

3.1
-0.03
1.2
0.2
-0.7
-1.45
2.2
-0.05
0.93
0.21

Table 4: Using Eq. 4 and 5, find estimates of μ and σ for this data.

```
x = [...];
[b,a] = hist(x,4); %(How many bins?)
```

Some vocabulary:

- **normalization** - Total probability is 1.0. For pdf - the integral from $[-\infty, \infty]$ is 1.0. For pmf - the sum from $[-\infty, \infty]$ is 1.0.
- **pmf** - Probability mass function: A *normalized continuous function* that gives the probability of a value, given the value.
- **histogram** - Histograms are an attempted measurement of the pmf by breaking the data into discrete bins. Histograms can be *normalized* as well.
- **pdf** - Probability density function: A *normalized continuous function* that gives the probability density of a value, given the value. Integrating the *normalized* pdf between two values gives the probability of observing data between the given values.

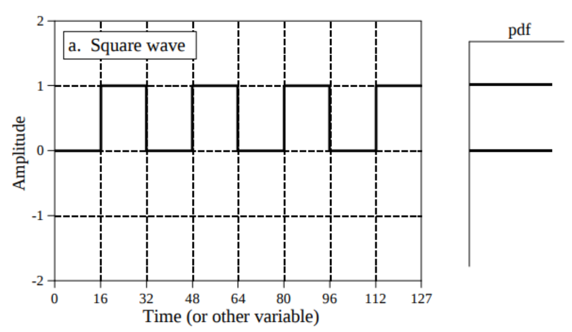

Figure 8: The square-wave signal spends equal time at 0.0 and 1.0, and the probability density function reflects that.

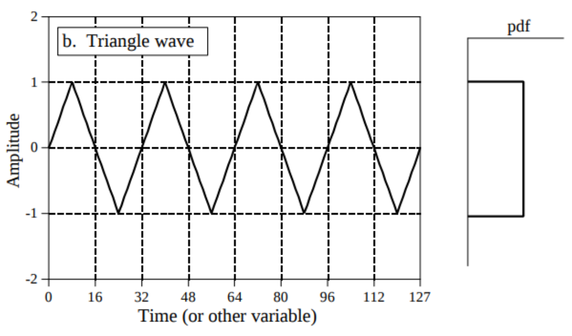

Figure 9: The triangle-wave signal spends equal time at all values *between* 0.0 and 1.0, and the probability density function reflects that.

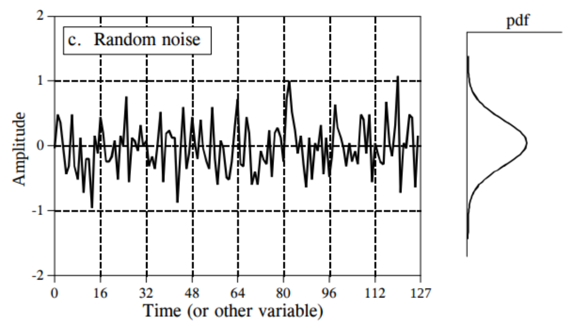

Figure 10: The random noise *usually* spends time near 0.0, but rarely it fluctuates to larger values.

Normal distribution

Normally distributed data decreases in probability at a rate that is proportional (1) to the *distance from the mean*, and that is proportional (2) to the *probability itself*.

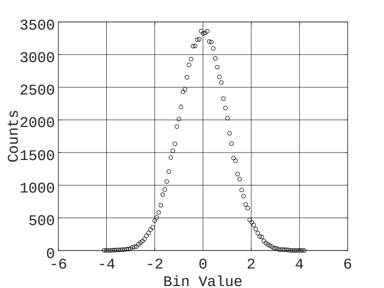

Figure 11: Normally distributed data counts decrease as measured further from the mean for *two reasons*.

Normal Distribution PDF

Let $p(x)$ be the PDF of normally distributed data x with mean μ. In order to obey conditions (1) and (2), the function $p(x)$ must be described by the following differential equation, where k is some constant.

$$\frac{dp}{dx} = -k(x - \mu)p(x) \tag{6}$$

Normal distribution

Rearranging Eq. 6, we have

$$\frac{dp}{p} = -k(x-\mu)dx \tag{7}$$

Integrating both sides gives

$$\ln(p) = -\frac{1}{2}k(x-\mu)^2 + C_0 \tag{8}$$

Exponentiating,

$$p(x) = C_1 \exp\left(-\frac{1}{2}k(x-\mu)^2\right) \tag{9}$$

Ensuring that the PDF is *normalized* requires

$$\int_{-\infty}^{\infty} p(x)dx = 1 \tag{10}$$

But how do we integrate Eq. 9? First, a change of variables. Let $s = \sqrt{k/2}(x - \mu)$, so $ds = \sqrt{k/2}dx$. Then, we have

$$C_1\sqrt{\frac{2}{k}}\int_{-\infty}^{\infty} \exp(-s^2)ds = 1 \tag{11}$$

Squaring both sides, we have

$$C_1^2\frac{2}{k}\left(\int_{-\infty}^{\infty} \exp(-s^2)ds\right)^2 = 1 \tag{12}$$

Let's pretend the two factors of the integral involve different variables:

$$C_1^2 \frac{2}{k} \left(\int_{-\infty}^{\infty} \exp(-x^2) dx \right) \left(\int_{-\infty}^{\infty} \exp(-y^2) dy \right) = 1 \tag{13}$$

Now we have

$$C_1^2 \frac{2}{k} \int_{-\infty}^{\infty} \exp(-(x^2 + y^2)) dx dy = 1 \tag{14}$$

Change to polar coordinates ($x^2 + y^2 = r^2$)

$$C_1^2 \frac{2}{k} \int_0^{\infty} \int_0^{2\pi} r \exp(-r^2) dr d\phi = 1 \tag{15}$$

Normal distribution

One more substitution: $u = r^2$, and $du = 2rdr$:

$$-\frac{C_1^2}{k}\int_0^\infty\int_0^{2\pi}\exp(-u)dud\phi = 1 \tag{16}$$

Solving for C_1, we find

$$C_1 = \sqrt{\frac{k}{2\pi}} \tag{17}$$

Thus the pdf of normally distributed data is

$$p(x) = \sqrt{\frac{k}{2\pi}}\exp\left(-\frac{1}{2}k(x-\mu)^2\right) \tag{18}$$

Let's defined $k = \frac{1}{\sigma_x^2}$ so that it's clear the exponent has the proper ratio of units:

$$\boxed{p(x) = \sqrt{\frac{1}{2\pi\sigma_x^2}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma_s}\right)^2\right)} \tag{19}$$

Statistics and Probability: Programming with Octave

More on the *hist* function in octave[2]

```
pkg install -forge io
pkg install -forge statistics
pkg load statistics
pkg help histfit
histfit(randn(1000,1))
histfit(rand(1000,1))
```

Let's work out the σ of a *flat* distribution between $[0, 1]$. What is it for a flat distribution between $[-1, 1]$? (We can derive this by hand as well if we cannot access statistics package).

[2] I hope this works, but if not, it's ok.

Some interesting notation for normal distributions:

$$N(\mu, \sigma) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{20}$$

Let's write a function **NGaus.m** that produces the Gaussian probability given μ and σ:

```
function ret = NGaus(mu,sigma,x)
    ...
endfunction
```

Now let's write a function *NRand* that sums N uniformly-distributed (flat) random variables x:

```
function ret = NRand(n)
    ret = sum(rand(n,1));
endfunction
```

Create a histogram of a few hundred outputs of *NRand*. What do you notice about the pmf? Let's plot *NGaus* on the same axes as the histogram of *NRand*. How do they compare?

We are on our way to producing $N(0,1)$ distributed numbers, and therefore our first **noise** signals...

The Box-Muller method for $N(0,1)$ distruted numbers:

$$X_1 = \sqrt{-2\ln(U)}\cos(2\pi V) \tag{21}$$

$$X_2 = \sqrt{-2\ln(U)}\sin(2\pi V) \tag{22}$$

Try this in octave... More vocabulary:

- **cdf** - Cumulative distribution function: Probability that a continuous random variable X is less than some value x. For a given pdf, the cdf $\Phi(X)$ is the integral of the total probability on $[-\infty, x]$. The derivative of the pdf is related to the pdf via the fundamental theorem of calculus.

If the pdf follows $f(x)$, then

$$\Phi(X \leq x) = \int_{-\infty}^{x} f(x)dx \tag{23}$$

The cdf of $N(0,1)$ has an expected shape, but can't be expressed with elementary functions.

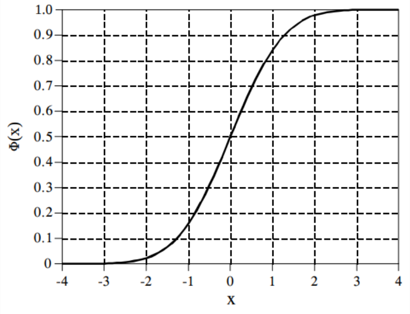

Figure 12: The cumulative distribution of the normal distribution. Although we can plot it, it's hard to write. We will discuss the *erf* and *erfc* functions in the near future.