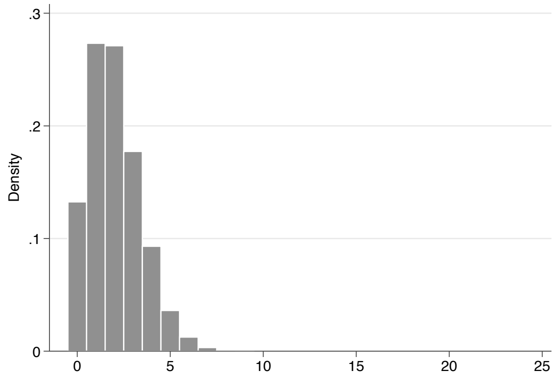
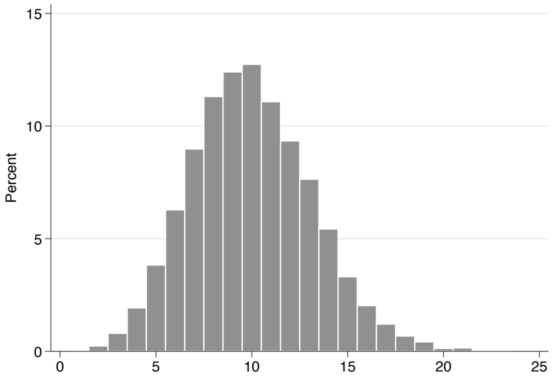
Activity 13: Point Estimates & Maximum Likelihood Estimation

In activities 8-11, we learned about the distributions of several discrete and continuous random variables. The location and shape of these distributions depended on some parameters. For example:

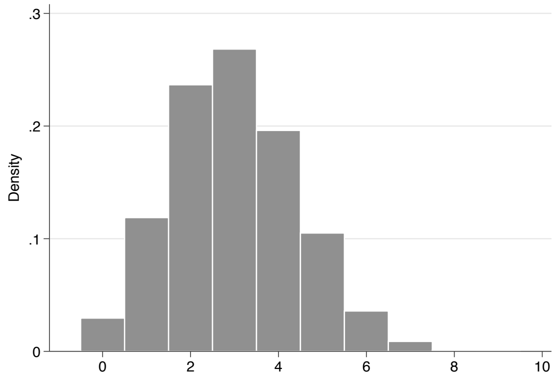
Poisson with λ=2



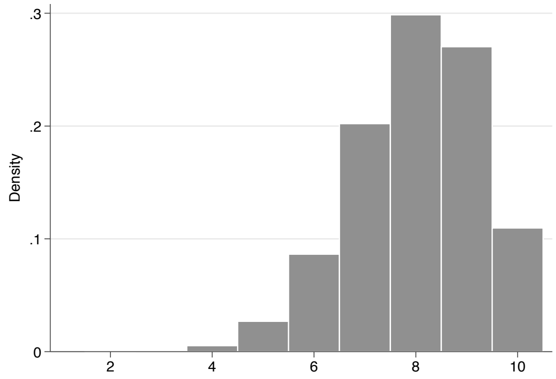
Poisson with λ=10



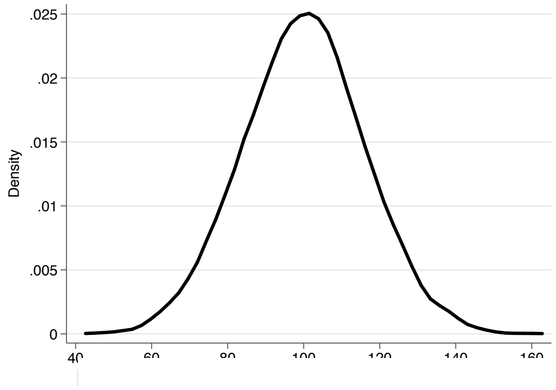
Binomial with p=0.30 and n=10



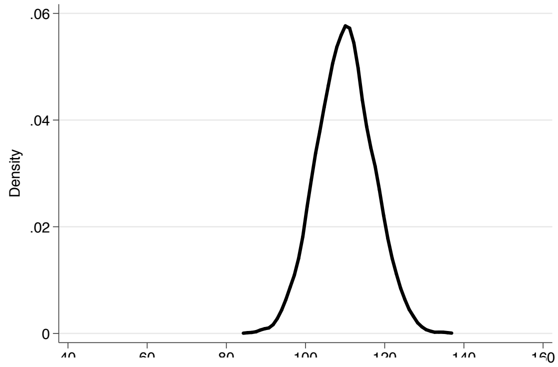
Binomial with p=0.80 and n=10



Normal with μ=100 and σ=16



Normal with μ=110 and σ=7



In order to use these distributions, we need to find reasonable estimates for the values of parameters we will never know (λ, p, μ, σ). For example, in a Poisson distribution, we've estimated λ as the average of our sample data. Likewise, in a Binomial distribution, we've estimated p by using the proportion of successes in our sample. In a normal distribution, we've estimated μ and σ by using the mean and standard deviation from our sample.

Are we justified in using these estimates? Are our estimates reasonable estimators of the parameters?

Maximum Likelihood Estimation

Scenario: Suppose we're interested in the proportion of SAU engineering students who find jobs within 1 year of graduation. We survey 20 alumni and learn 9 of them found a job within a year.

The actual data from the survey are: **1 0 0 0 1 1 0 1 0 0 1 0 1 1 0 1 0 1 0 0**
(where 1 = found a job and 0 = did not find a job)

We're interested in finding Φ = the proportion of **all** engineering students who find a job within a year of graduation. Notice that we'll never know the true value of this parameter.

1. It looks like we're interested in a Binomial Distribution.
   We have a series of independent trials (students)
   Each trial has two possible outcomes (finding a job or not)
   We have a constant probability of success (finding a job); we just don't know the value of that probability.
   Using our sample data, what's our best estimate of Φ?

2. To make things look more familiar, we'll use *p* to represent the probability of finding a job. If a binomial distribution finds this scenario, the likelihood of observing exactly 9 students who find jobs out of 20 total students would be:

$$P(X=9)=\binom{20}{9}p^{9}\left(1-p\right)^{20-9}=\left(167960\right)p^{9}\left(1-p\right)^{11}$$

We want to find the "best" estimate of *p*. Because we actually did observe 9 successes out of 20 trials, it seems reasonable to stipulate that the best estimate of *p* would be the one that makes this likelihood as large as possible.

In other words, **the best estimate of *p* would be one that maximizes the likelihood of observing our data**.

Since we know 0 ≤ p ≤ 1, we could use guess-and-check to find the best estimate of *p*. For example, I could arbitrarily guess values of 0.10, 0.50, and 0.80:

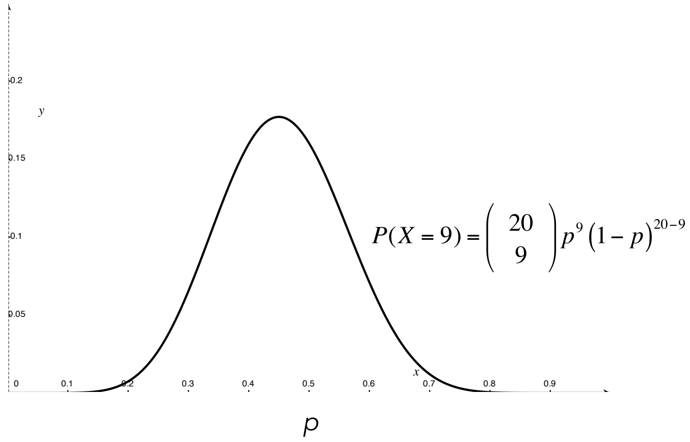
$$P(X=9|p=0.10)=\left(167960\right)\left(0.10\right)^{9}\left(1-0.10\right)^{11}=0.0000527$$

$$P(X=9|p=0.50)=\left(167960\right)\left(0.50\right)^{9}\left(1-0.50\right)^{11}=0.1601791$$

$$P(X=9|p=0.80)=\left(167960\right)\left(0.80\right)^{9}\left(1-0.80\right)^{11}=0.0004617$$

From these guesses, which value of *p* appears to be our best estimate of the true parameter, Φ?

3. Guess-and-check could lead us to a good estimate, but it would be an inefficient way to search for the best estimate. Since we want to maximize our likelihood function, let's graph it and see if we can visually find the maximum.



It looks like the maximum is located somewhere between 0.40 and 0.50. To find the actual maximum, we can set the first derivative of our likelihood function equal to zero and solve for *p*:

$$\frac{d}{dp}\left[\binom{20}{9}p^{9}\left(1-p\right)^{20-9}\right]$$

The combination (20 choose 9) is simply a constant across all values of *p*, so we can ignore it. If we took the derivative, we'd need to use the product rule. To eliminate the products, we can take the natural logarithm of this function before we differentiate:

$$\ln\left[p^{9}\left(1-p\right)^{20-9}\right]=(9)\ln(p)+(11)\ln(1-p)$$

$$\frac{d}{dp}\left[(9)\ln(p)+(11)\ln(1-p)\right]=0$$

$$\frac{9(1-p)+11(p)}{p(1-p)}=\frac{9-9p+11p}{p-p^{2}}=\frac{9-20p}{p-p^{2}}=0$$

$$9-20p=0$$

$$p=\frac{9}{20}=0.45$$

From this, we've found the best estimate of *p* would be 0.45. In other words, 0.45 is the value of *p* that maximizes the likelihood that we would have observed the data that we actually observed.

It just so happens that 0.45 = 9/20 (the proportion of successes in our sample of data). Does this result always hold for the binomial distribution? If we want the best estimate of Φ, can we simply use the proportion of successes from our sample of data?

To find out, we can generalize our calculations. Let's imagine a binomial random variable with *n* trials and probability of success = *p*...

To find the value of *p* that maximizes the probability (or likelihood) of observing any data that we observe from this random variable, we would do the following:

$$P(X=x)=\begin{pmatrix}\text{n}\\ x\end{pmatrix}p^{x}\left(1-p\right)^{n-x} \qquad \text{<-- write out our likelihood model}$$

$$\ln\left[p^{x}\left(1-p\right)^{n-x}\right]=(x)\ln(p)+(n-x)\ln(1-p) \qquad \text{<-- take the log to make differentiation easier}$$

$$\frac{d}{dp}\left[(x)\ln(p)+(n-x)\ln(1-p)\right]=0 \qquad \text{<-- set the derivative equal to zero to maximize}$$

$$\frac{x}{p}-\frac{n-x}{1-p}=0 \qquad \text{<-- take the derivative}$$

$$\frac{x}{p}=\frac{n-x}{1-p} \qquad \text{<-- solve for p}$$

$$x(1-p)=(n-x)p$$

$$x-px=np-xp$$

$$x=np$$

$$p=\frac{x}{n} \qquad \text{<-- Find the maximum likelihood estimate of p}$$

This demonstrates that the best estimate of Φ is simply use the proportion of successes from our sample of data

Suppose we have random variable X with probability function f(x;θ). The probability function could be anything (e.g., binomial, poisson, exponential; or any other discrete or continuous distribution); θ is a parameter of the function we wish to estimate.

Let x_1, x_2, ... , x_n represent a sample of n independent observed values of X.

The **likelihood function** is the product of the probability function estimate at each observed value:

$$L\left(\theta\right)=\prod_{i=1}^{n}f\left(x_i;\theta\right)$$

The **maximum likelihood estimate (MLE)** of θ is the value that maximizes L(θ) for all possible values of θ.

It's often easier to maximize the natural logarithm of the likelihood function, as we'll see in the next few examples.

Scenario: Using something called *item response theory*, we can model the probability that an examinee answers a question correctly on a test. This probability depends on:

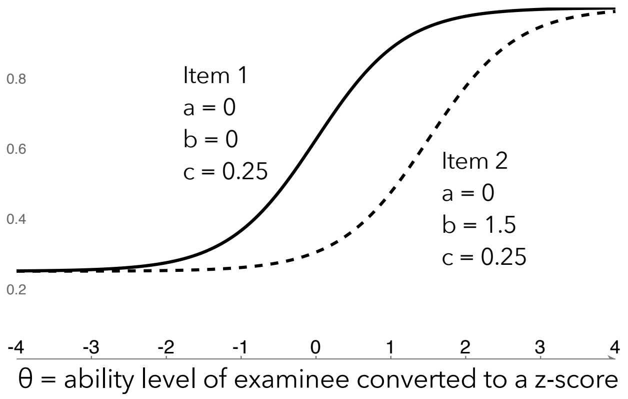
- the ability level of the examinee (which we will denote as θ)
- the difficulty of the question (which we will denote as b)
- the probability that a student can guess the answer (which we will denote as c)
- how well the item discriminates between students who know or do not know the content (a)

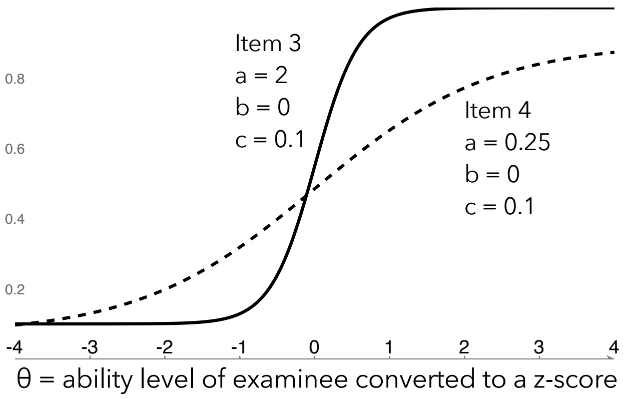
With these parameters, we can model the probability using a logistic function:

$$P\left(\text{correct} \mid \theta, \text{a}, \text{b}, \text{c}\right) = c + (1-c)\frac{e^{1.7a(\theta-b)}}{1+e^{1.7a(\theta-b)}}$$

Don't worry about how this function was derived. I just wanted to show an application of MLE.

4. To give you an idea of what this probability model looks like, I've graphed 4 different *item characteristic curves* below. They all use the logistic function, just with different values for the parameters a, b, and c.





Look at the two curves on the left. These curves only differ in terms of their b parameters. Which question, Item 1 or Item 2, appears to be more difficult?

Look at the two curves on the right. These curves only differ in terms of their a parameters. Explain what the *a* parameter represents.

The two curves on the right have c = 0.25. The two curves on the left have c = 0.10. Explain what the c parameter represents.

5. Suppose we give a 4-question test to a student. The parameters for the 4 questions are found to be*:

	a	b	c
Question 1	1 / 1.7	-1	0.15
Question 2	1 / 1.7	0	0.15
Question 3	1 / 1.7	1	0.25
Question 4	1 / 1.7	2	0.15

Suppose the student only answers questions 1 and 3 correctly. In other words, we have 4 observed data points of 1, 0, 1, 0. How can we best estimate the ability level of this student?

To find the MLE, we must maximize the likelihood function: $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$

where $f(x) = c + (1-c)\frac{e^{1.7a(\theta-b)}}{1+e^{1.7a(\theta-b)}}$

Let's first write out the likelihood function for each test question by plugging-in the item parameters:

Q #1: $f(x_1) = 0.15 + (1-0.15)\frac{e^{(\theta+1)}}{1+e^{(\theta+1)}}$

Q #2: $f(x_2) = 1 - \left[0.15 + (1-0.15)\frac{e^{(\theta)}}{1+e^{(\theta)}}\right]$

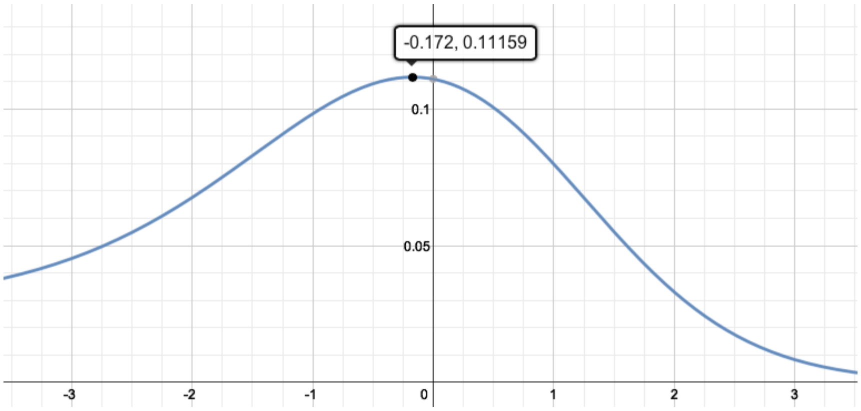
Q #3: $f(x_3) = 0.25 + (1-0.25)\frac{e^{(\theta-1)}}{1+e^{(\theta-1)}}$

Q #4: $f(x_4) = 1 - \left[0.15 + (1-0.15)\frac{e^{(\theta-2)}}{1+e^{(\theta-2)}}\right]$

If we multiply these together, we get the likelihood of observing a student who scores 1, 0, 1, 0 on the questions.

$$L(\theta) = \left(0.15 + (1-0.15)\frac{e^{(\theta+1)}}{1+e^{(\theta+1)}}\right)\left(1 - \left[0.15 + (1-0.15)\frac{e^{(\theta)}}{1+e^{(\theta)}}\right]\right)\left(0.25 + (1-0.25)\frac{e^{(\theta-1)}}{1+e^{(\theta-1)}}\right)\left(1 - \left[0.15 + (1-0.15)\frac{e^{(\theta-2)}}{1+e^{(\theta-2)}}\right]\right)$$

To find the MLE, we need to find the value of θ that maximizes this function. Because this is a complicated example, I'll find the maximum graphically:



It looks as though our best estimate of this examinee's ability is -0.172 (about 0.2 standard deviations below the mean).

* How do we estimate these item parameters? We can use something called joint maximum likelihood estimation

6. Suppose we're interested in modeling the time we wait to checkout at a grocery store as an exponential distribution. From activity #10, we know the probability function for an exponential distribution is of the form: $f(x) = \lambda e^{-\lambda x}$ for some value of λ>0.

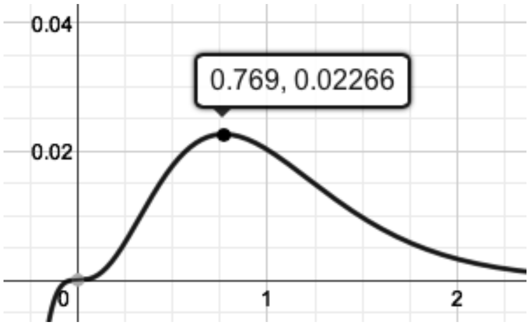
   Further, suppose we go to this grocery store 3 times and find we wait 0.4, 1.1, and 2.4 minutes. The average of these three waiting times is equal to 1.3 minutes.

   From this information, what's your best estimate for λ in this scenario?

   Let's find the MLE for λ by first writing out and simplifying the likelihood function.

$$L(\theta) = \left(\lambda e^{-\lambda(0.4)}\right)\left(\lambda e^{-\lambda(1.1)}\right)\left(\lambda e^{-\lambda(2.4)}\right) = \lambda\lambda\lambda e^{-0.4\lambda} e^{-1.1\lambda} e^{-2.4\lambda} = \lambda^3 e^{-3.9\lambda}$$

   We can now maximize it graphically or by taking the first derivative of the natural log of the likelihood function.



$$\ln\left[L(\theta)\right] = \ln\lambda^3 + \ln\left(e^{-3.9}\right) = 3\ln\lambda - 3.9\lambda$$

$$\frac{d}{d\lambda}\left[3\ln\lambda - 3.9\lambda\right] = \frac{3}{\lambda} - 3.9$$

   Setting this derivative equal to zero, we find λ = 1/1.3 = 0.769

Point Estimate: We symbolize an unknown population parameter as θ (theta).

A point estimate of theta is denoted as $\hat{\theta}$ (theta circumflex or theta hat).

Scenario: Suppose I'm interested in certain parameters about books in our library. I want to know these parameters for ALL books in the library – even books that our library doesn't have yet – so I'll never be able to know these parameter values.

I can take a sample of books and calculate some estimates of these parameters. I'm interested in determining which estimates would be best for the parameters I'm interested in.

Parameters that I'd like to estimate	Possible estimators from a sample of books
1. The average weight of all books: $\theta = \mu$	1a. The average weight: $\hat{\theta} = \bar{X} = \frac{1}{n}\sum x$ 1b. The trimmed mean (after eliminating 10% of the tails) 1c. The median weight
2. The standard deviation of book weights, $\theta = \sigma$	2a. The standard deviation: $\hat{\theta} = \hat{\sigma} = \sqrt{\frac{\sum (x - \bar{X})^2}{n}}$ 2b. The standard deviation: $\hat{\theta} = s = \sqrt{\frac{\sum (x - \bar{X})^2}{n-1}}$
3. The proportion of books that are less than five years old: $\theta = \Phi$	3a. The proportion <5 years old: $\hat{\theta} = p = \frac{\text{\# less than 5 years old}}{\text{books sampled}}$
4. The biggest ISBN number of all books currently in the library	4a. The biggest ISBN number in our sample 4b. $\hat{\theta} = \left(\text{biggest ISBN - 1}\right) + \frac{\text{biggest ISBN}}{\text{\# of books sampled}}$

Notice that population parameters are symbolized with greek letters while sample statistics use english letters

7. Which estimator would you choose as the best estimate for each parameter value?

The best estimator for #1 (average weight of all books) is: ______________________________

The best estimator for #2 (std. deviation of the weight of all books): ______________________________

The best estimator for #4 (biggest ISBN number): ______________________________

We're going to need some criteria for choosing the best estimate of parameter values.
What are the characteristics of good estimators?

Good estimates are unbiased

We want $E\left[\hat{\theta}\right]=\theta$. Bias is represented by $\text{Bias}=E\left[\hat{\theta}\right]-\theta$

8. Suppose we repeatedly take samples of size n=25 from an unknown population and calculate $\hat{\theta}$ for each sample. If our estimator is unbiased, what would we want to be true about the estimates we calculate from each sample?

9. Suppose we're interested in the average weight of all books in the library. We sample n=25 books and calculate the average weight. Is this sample mean an unbiased estimator of the population mean? Does $E\left[\bar{X}\right]=\mu$?

10. Does this mean that our sample average is a good estimate of the population average?

11. Suppose we're interested in the proportion of books that are less than 5 years old. We sample n=25 books and calculate the proportion that are less than 5 years old. Is this an unbiased estimator of the population proportion? Recall that the expected value of a binomial random variable is E[x] = np.

12. Suppose we're interested in the standard deviation of the weights of all books in the library. We sample n=25 books and calculate... well, what should we calculate? Should we calculate a standard deviation with *n* or *n-1* in the denominator? Why?

Good estimates have small variances.

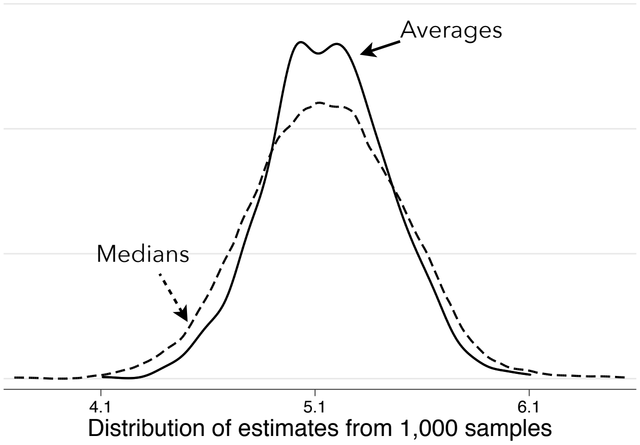
We want $\mathrm{Var}\left[\hat{\theta}\right]$ to be small (at least smaller than what we'd get from other potential estimators).

13. What does this mean? Well, even if an estimate is unbiased, it might fluctuate wildly from sample to sample.

    If we're going to estimate a population parameter, we're probably going to take a single sample and calculate a single unbiased estimate for that parameter. We would hope the value of our estimate is close to the value of the parameter.

    Suppose we're interested in the average weight of all books in the library. We take a sample of n=25 books and calculate the average age. We then take another sample and calculate another average. We do this again and again until we have 1,000 sample averages.

    At the same time, we calculate the median for each of our 1,000 samples. We then graph all 1,000 averages and medians we calculated from our samples:



    Given that we would only take one sample and calculate a single estimate, would you choose to use the sample average or sample median as the best estimate of the population average? Why?

14. Calculate the variance of the sample average (mean). What's the standard deviation of the sample mean? We call this the *standard error*.

15. What does this tell us to do if we want to get a more accurate estimate of the population mean?

16. So far, we've learned a few things:

    - If we want to estimate μ, we should use the sample mean. Why? Because it's unbiased and we can calculate its variance. It's also the maximum likelihood estimate of the population mean.

    - If we want to estimate the population proportion, we should use the sample proportion. Why? Because it's unbiased and it's also the MLE.

    We still don't know what estimator to use if we want to estimate the population standard deviation. In the next few pages, I'm going to try to convince you that:

    - The best estimate of the standard deviation has n-1 in the denominator: $\hat{\sigma} = s = \sqrt{\dfrac{\sum\left(x-\bar{X}\right)^2}{n-1}}$

    If this estimate is unbiased, then $E\left[s\right] = \sigma$. To show this, let's simplify the numerator of our estimator:

$$\sum(x_i-\bar{X})^2 = \sum\left[(x_i-\mu)-(\bar{X}-\mu)\right]^2 = \sum\left\{\left[(x-\mu)-(\bar{X}-\mu)\right]\left[(x-\mu)-(\bar{X}-\mu)\right]\right\} =$$

$$= \sum\left[(x-\mu)^2 + 2(x-\mu)(\bar{X}-\mu) + (\bar{X}-\mu)^2\right] =$$

$$= \sum(x-\mu)^2 - 2(\bar{X}-\mu)\sum(x-\mu) + \sum(\bar{X}-\mu)^2$$

$$= \sum(x-\mu)^2 - 2(\bar{X}-\mu)\left[\sum x - \sum\mu\right] + n(\bar{X}-\mu)^2 =$$

$$= \sum(x-\mu)^2 - 2(\bar{X}-\mu)\left[n\bar{X} - n\mu\right] + n(\bar{X}-\mu)^2 =$$

$$= \sum(x-\mu)^2 - 2n(\bar{X}-\mu)^2 + n(\bar{X}-\mu)^2 = \sum(x_i-\mu)^2 - n(\bar{X}-\mu)^2$$

If s² is unbiased, then $E\left[s^2\right] = E\left[\dfrac{\sum\left(x-\bar{X}\right)^2}{n-1}\right] = \sigma$, or equivalently, $E\left[s^2\right] = E\left[\sum\left(x-\bar{X}\right)^2\right] = (n-1)\sigma$

$$E\left(\sum(x-\bar{X})^2\right) =$$

$$= E\left[\sum(x_i-\mu)^2 - n(\bar{X}-\mu)^2\right] = \sum\left[E(x_i-\mu)^2\right] - E\left[n(\bar{X}-\mu)^2\right] = \sum\sigma^2 - nE(\bar{X}-\mu)^2 = \sum\sigma^2 - nVar(\bar{X}) = n\sigma^2 - n\left(\frac{\sigma^2}{n}\right) =$$

$$= n\sigma^2 - \sigma^2 = \sigma^2(n-1)$$

$$E(s^2) = E\left(\frac{\sum(x-\bar{X})^2}{n-1}\right) = \frac{1}{n-1}E\left[\sum(x-\bar{X})^2\right] = \frac{1}{n-1}\left[\sigma^2(n-1)\right] = \sigma^2$$

This demonstrates that our estimator is unbiased.

17. If, for some strange reason, you didn't like that proof, here's another attempt at an explanation.

Recall the definition of the sample mean we found in activity #12. We said it's the value that minimizes the sum of squared deviations within a sample of observations.

Suppose we have three data points: 1, 2, 6. We want to find a number that best represents the center of this data. We'll arbitrarily choose 2 and 3 as our best guesses for the center of the data.

	Trying 2 as the best center			Trying 3 as the best center		
X	c = center	X - c	$(X - c)^2$	c = center	X - c	$(X - c)^2$
1	2	-1	1	3	-2	4
2	2	0	0	3	-1	1
6	2	4	16	3	3	9
			Sum = 17			Sum = 14

As we saw in activity #12, no matter what number we choose for *c*, we'll never get a sum less than 14.

Now consider a situation in which we don't know the value of μ. We'll use our best estimate, the sample mean, but it won't be exactly the same as μ.

Which of the following will be larger and which of the following will be the "true" value of the population standard deviation?

$$\hat{\sigma} = \sqrt{\frac{\sum (x-\mu)^2}{n}} \qquad \hat{\sigma} = \sqrt{\frac{\sum \left(x-\bar{X}\right)^2}{n}}$$

The larger one is: ______________ ______________

The "correct" one is: ______________ ______________

The one we can calculate: ______________ ______________

What's the take away from all of this?

18. If we have a computer that's able to run a java applet, go to the following website:
http://www.onlinestatbook.com/stat_sim/sampling_dist/index.html