[Print: May 31, 2017]
[]

1 Statistical Goodies

1.1 Nomenclature.

1.1.1 Population and Sample.

Population. *A population is the set of everything that meets a specific criteria.*

Here are some examples of populations:

- all women who are US citizens,
- all the money in my wallet (might be the empty set)
- the eight achievable results produced by tossing a coin three times,
- all results achieved by repeatedly tossing a coin three times,
- the AMTH 108 class.

Random Sampling. *A random sampling from a population is a set of independent observations taken of a population. The term can also refer to an independent set of representatives taken from a population.*

Here are some examples of a random sampling:

- all women who are currently US citizens taken from the population of all women
- all the money in my wallet (might be the empty set) representing the population of money in my wallet for the month
- all results achieved by repeatedly tossing a coin three times representing the population of the eight achievable results produced by tossing a coin three times,
- the AMTH 108 class representing the population of university students.

Here is an example. The distribution of heights of everyone in the AMTH 108 class (considered as a population) is completely known, possess a density and cumulative distribution function, has a mean, a variance, and a moment generator. However, as a random sampling, representing (say) the heights of everyone in at the university, the density function, distribution, etc. of the sample may or may not accurately represent the density, distribution, etc. of the population of university students, or of everyone in the county, or of everyone in the state, etc.

The job of statistics is to determine how accurately the probabilistic properties of a random sampling represent the probabilitstic properties of a larger population.

1.2 Random Variables and Samples.

1.2.1 Basic Stuff. Let X be a random variable defined on some population. If an individual, i, is chosen at random from the population, then the value taken by X on i is a random variable in its own right called X_i.

The random variables, X_i, are all drawn from the same overarching distribution, X, and consequently each possess the same distribution as X. Also if the selection processes of the individuals is truly random, each of the X_i as a random variable will be independent of all other X_j in the sample.

A Random Sample. *A random sample (or just a sample) is a collection, usually finite in number, of n independent and identically distributed (iid) random variables.*

Note that a sample is *not* a collection of numbers but rather a collection of random variables. Usually, each of these random variables is expressed, that is, given a value, $\widetilde{X}_i$, usually via measurement or by running some experiment.

1.2.2 Histograms. Given a sample $X_i, i = 1, 2, \ldots, n$ taken from some master process X and the associated set of sample values $\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_n$, a histogram is an approximation to the density function for the over-arching master process X (See figure on following page).

1. To construct this approximation, begin by choosing an interval $[a, b]$, lying on the X axis, that contains all of the expressed values. This is another way of saying that each value $\widetilde{X}_i$ should fall between a and b (i.e. $a \leq \widetilde{X}_i \leq b$).
2. Divide the interval $[a, b]$ into n bins, labelled $0, 1, 2, \ldots, n-1$, of width w. Do this by selecting either n or w. The relationship between a, b, w, and n is $wn = b - a$. Note that either w or n may be chosen at this juncture and the other parameter (that is, n or w) determined from the basic relationship. Note, however, that while n may be chosen arbitrarily and w determined from it, the opposite is not generally true. That is, selecting w must be done in such a way that $(b - a)/w$ is an integer. For example, the choices $b = 11, a = 0, w = 2$ are incompatable because $n = (b - a)/w = (11-0)/2 = 5.5$, which is not an integer. However, these nearby choices are compatible: $(b = 12, a = 0, w = 2)$, $(b = 11.5, a = -0.5, w = 1)$, $(b = 11, a = -1, w = 3)$, and so on. Often, when the values $\widetilde{X}_i$ are all integers it is desireable to have $a = \min(\widetilde{X}_i) - \frac{1}{2}, b = \max(\widetilde{X}_i) + \frac{1}{2}, w = 1$.
3. Each data point $\widetilde{X}_i$ may be represented as follows:
$$\widetilde{X}_i = a + kw + \epsilon w,$$
where a is the left edge of the interval, w is the bin width, k is number of whole bins between a and $\widetilde{X}_i$, and $0 \leq \epsilon w \leq w$ with $0 \leq \epsilon \leq 1$ is a fraction of a bin. Dividing both sides of this equation by w yields:
$$\frac{\widetilde{X}_i - a}{w} = k + \epsilon.$$
Since $0 \leq \epsilon \leq 1$ and k is an integer it must be that
$$k = \left[\frac{\widetilde{X}_i - a}{w}\right].$$

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Min Samp:	0.453084419	a	0.45	h :	0.001347089								
2	Max Samp:	18.26992754	b	19	w :	1.855								
3	Num Samp:	400	n	10	Max bin num :	9								
4														
5	Data Chi-Sq 5 d.f.	Bins	Bin Number	Histogram	Histogram Normalized to Unit area	Theoretical Representatives	Theoretical Distribution							
6	4.28385383	2	0	78	0.105121938	0.45	0.032054671							
7	1.875855098	0	1	106	0.142857429	2.805	0.153670356							
8	3.255970741	1	2	105	0.14150434	5.16	0.118109165							
9	1.40850988	0	3	45	0.060646003	7.515	0.063947088							
10	7.081130978	3	4	35	0.047169813	9.87	0.029649766							
11	2.77042978	1	5	13	0.017202156	12.225	0.012590329							
12	9.748993466	5	6	6	0.008086234	14.58	0.005051476							
13	4.852682677	2	7	7	0.009439623	16.935	0.001947951							
14	6.383521074	3	8	2	0.002695417	19.29	0.000729486							
15	7.745852283	3	9	3	0.004031267	21.645	0.0002671							
16	5.987683096	2	10	0	0	24	9.6066E-005							
17	6.353925812	3	11	0	0	26.355	0.000034054							
18	3.056853224	1	12	0	0	28.71	1.1927E-005							
19	5.281669145	2	13	0	0	31.065	4.1354E-006							
20	10.84641568	5	14	0	0	33.42	1.4215E-006							
21	1.669303841	0	15	0	0	35.775	0.000000485							
22	1.679819599	0	16	0	0	38.13	1.6439E-007							
23	3.706656533	1	17	0	0	40.485	5.5402E-008							
24	3.59102036	1	18	0	0	42.84	1.8577E-008							
25	5.62844764	2	19	0	0	45.195	6.2009E-009							
26	3.065203892	1	20	0	0	47.55	2.0614E-009							
27	4.187756766	2	21	0	0	49.905	6.8277E-010							
28	2.917256654	1	22	0	0	52.26	2.2539E-010							
29	5.816879286	2	23	0	0	54.615	7.4175E-011							
30	2.607915607	1	24	0	0	56.97	2.4343E-011							
31	4.480381307	2	25	0	0	59.325	7.9687E-012							
32	2.835003838	1												
33	2.16206232	0												
34	0.510200829	0												
35	5.313105601	2												
36	3.521944062	1												
37	4.343746118	2												
38	6.687754243	3												
39	6.27184304	3												
40	4.930362758	2												
41	4.653643302	2												

Since the bins are numbered beginning with 0, bin k is the bin into which $\widetilde{X}_i$ falls naturally.

4. Each time $\widetilde{X}_i$ falls into some bin k ($0 \leq k < n$) according to the rule in item 3, place a rectangle over the bin base (on the horizontal axis) that has height h and width w (the same as the bin width). When completed there will be n_k of these rectangles placed in each bin forming a larger (and usually vertically elongated) rectangle of area $n_k wh$. The total area under all such rectangles is

$$A = \sum_{k=0}^{n-1} whn_k = wh \sum_{k=0}^{n-1} n_k = whN,$$

where N is the total number of sample points.

5. Complete the histogram by normalizing the height, h, of each component rectangle so that $A = 1$. That is, set the value of h to be

$$h = \frac{1}{wN}.$$

Because the theoretical density function for a distribution traps an area equal to 1 underneath it's "curve" and above the x-axis, a histogram will not resemble a density curve at all if the area under the histogram isn't also normalized to 1. (Note: for Distrete random variables, the "curve" is just a set of discrete points lying above the x-axis.)

To produce a side-by-side comparison of a histogram with a theoretical density, first select one point within each bin to represent the bin (this point is completely arbitrary and *doesn't* have to be one of your data values, but a good practice is to locate this point in the same relative location within each bin). Call this point the bin representative, r_i. The bin boundaries are given by the formula:

$$b_i = a + iw \qquad i = 0, 1, 2, \ldots, n,$$

so the bin representatives satisfy $b_i \leq r_i \leq b_{i+1}$. Note that this definition permits a bin representative to be on either the left or right edge of a bin (but refrain from selecting the common boundary of two adjacent bins as the bin representative of each bin). Often, a good choice for the representive is the midpoint of the bin: $b_i + w/2$ or $i = 0, 1, 2, \ldots, n-1$.

A theoretical density function, $f_X(x)$, has one nominal input, usually called x. However, to specify an exact density function may require the estimation of many parameters involved with the density. For example, to specify an exact Normal distribution, the mean μ and the variance σ^2 must be specified; to specify an exact Binomial distribution, the parameters n and p must be specified; and so on. Once estimates for the parameters have been calculated (see the next section on estimators and estimates), the theoretical density may be plotted by assigning the variable x the values of the bin representatives, r_i. If sufficiently many samples are at hand, there should be good agreement between these two charts.

One final thing to note is that there will almost always be poor agreement between the histogram and the theoretical distribution in the tails of the distribution. These are the

regions where the probability that X takes on these values is small. Consequently, it isn't likely that many such values will be encountered during the data gathering phase of any experiment. This under representation will distort the comparison between density function and histogram).

1.3 Popular Statistics. A statistic is a function defined on a sample of size n. That is, the statistic K is given by $K = f(X_1, X_2, \ldots, X_n)$. Here are the popular ones:

- The Sample Mean: $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$
- The Sample Variance: $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
- The Maximum: $M = \max(X_i)$ and the Minimum: $m = \min(X_i)$
- The Range: $R = M - m$.
- The Median: Sort X_i in increasing order. The median is $X_{n/2}$ if n (the number of samples) is even, and is $(X_i + X_{i+1})/2$ if n is odd where $i = (n-1)/2$.

1.3.1 Population Parameters. Quite frequently, a distribution will have parameters associated with it. Here are a few examples:

Distribution	**Parameters**
Geometric	p
Binomial	n and p
Poisson	k or λ
Gamma	α and β
Normal	μ and σ

In general a parameter that is associated with a distribution will be referred to in general by the Greek letter θ.

1.3.2 Estimators. One of the goals of statistics is to provide mechanisms for estimating the parameters found in various distributions. To do this, samples are taken from the distribution and combined to form a statistic called an estimator. The estimator for a parameter, θ, is usually denoted by placing a carat over the variable name: $\widehat{\theta} = f(X_1, X_2, \ldots, X_n)$. Once the sample is taken, the value that results by combining the sample values together via the estimator is call an *estimate* and is usually denoted by placing a tilde over the variable: $\widetilde{\theta} = f(x_1, x_2, \ldots, x_n)$ where $X_i = x_i$ are the values assumed by the sample variables.

What makes an estimator different from a statistic is that an estimator is a statistic that is targeted towards a specific parameter. Consider the mean, μ, of a distribution. This is a parameter of a distribution – although μ need not appear in the expressions for the density or distribution function. Nevertheless, it is still a parameter. The same is true of the variance. The most typical estimator used for the mean is the sample mean: $\widehat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Are there other estimators for the mean? Certainly! One could use $\widehat{\mu} =$ median or $\widehat{\mu} = (M + m)/2$ or ever $\widehat{\mu} = 12$. All are valid estimators that produce valid estimates when evaluated (some are better than others). Which are the good ones to use?

The qualities that create a good estimator for a parameter are these:

- The estimator is unbiased

- The variance of the estimator tends to zero and the number of samples increases without bound.

Remember that an estimator, being an statistic, is a random variable and so it makes sense to talk about the mean and variance of an estimator.

1.3.2.1 Biased v. Unbiased. An estimator $\widehat{\theta}$ (for θ) is unbiased if $E[\widehat{\theta}] = \theta$. The sample mean estimator, $\widehat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, is unbiased for the mean. So is the sample variance estimator, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, unbiased for the variance (see last page).

However $S = \sqrt{S^2}$, the sample standard deviation estimator, is a biased estimator of the standard deviation.

1.3.2.2 Behavior as Samples Increase. The second criteria for a good estimator is that the variance of the estimator must tend to zero as the number of samples goes to infinity. Consider the sample mean, $\bar{X}$. The variance of this statistic is

$$\mathrm{Var}(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n},$$

since the random variables X_i are independent and identically distributed . Note that this variance has the desired size property: $\lim_{n\to\infty}\mathrm{Var}(\overline{X}) = \lim_{n\to\infty}\frac{\sigma^2}{n} = 0$. The associated standard deviation of $\overline{X} = \sigma/\sqrt{n}$, has such widespread usage that it is called the standard error of the mean, sometimes denoted by σ_e.

1.3.3 Creating Estimators. A simple way of creating an estimator is to use the method of moments. The k-th moment is defined by $\mu_k = E[X^k]$. An estimator for this that seems like it might be a good choice is $\widehat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$. Just as for the sample mean, this estimator is unbiased and $\mathrm{Var}(\widehat{\mu}_k) \to 0$ as $n \to \infty$. Note that following the same logic used for the sample mean

$$\mathrm{Var}(\widehat{\mu}_k) = \frac{1}{n^2}\sum_{i}\mathrm{Var}(X_i^k) = \frac{\mathrm{Var}(X^k)}{n},$$

where X represents the over-arching process. The quantity $\mathrm{Var}(X^k)$ is the variance of the random variable X^k and is also the variance of all of the iid samples X_i^k. This is a constant depending only on the over-arching process and k.

In and of itself, these estimators may not be terribly interesting, however, a great many parameters associated with distributions may be expressed as arithmetic, polynomial, rational, etc. combinations of the moments of the over-arching process. Once the relationship that connects the moments to the parameter of interest is discovered, it is a simple matter to replace all of the variables with estimators. For example: the variance is related to the moments of a random variable by $\sigma^2 = \mu_2 - \mu_1^2$. Consequently, another estimator for the variance is $\widehat{\sigma}^2 = \widehat{\mu}_2 - (\widehat{\mu}_1)^2$. Note that this estimator may not be as good as S^2 (it is biased for one thing – see last page).

1.4 Confidence Intervals.

1.4.1 Generalities. Begin with a parameter, θ, about which a confidence interval is required. To find a confidence interval for this parameter, a form (a random variable associated with a parameter[1]), Y_θ, with a known distribution is needed that involves the parameter θ. The form chosen should have the property that it is relatively easy to move from an inequality of the sort $L_1 \leq Y_\theta \leq L_2$ to one of the sort $a \leq \theta \leq b$.

A confidence interval at the $1-\alpha$ level is defined to be an interval about θ such that $1-\alpha = P[a \leq \theta \leq b]$. To construct this interval begin by finding an interval around Y_θ satisfying $1-\beta = P[L_1 \leq Y_\theta \leq L_2]$. Since there are infinitely many values of L_1 and L_2 that satisfy this equation, restrict the choice by insisting that the amount of complementary probability (β) be evenly divided between the region $Y_\theta \leq L_1$ (the left tail) and $Y_\theta \geq L_2$ (the right tail). Determination of the values L_1 and L_2 requires the solution to these two tail-point problems:

$$\begin{aligned} P[Y_\theta \geq L_2] &= \beta/2 \qquad \text{(the right tail point)} \\ P[Y_\theta \geq L_1] &= 1-\beta/2 \quad \text{(the left tail point),} \end{aligned}$$

that is $L_1 = y_{\theta,1-\beta/2}$ and $L_2 = y_{\theta,\beta/2}$. By computing complementary probabilities, these two equations are the same as

$$\begin{aligned} F_{Y_\theta}(L_2) &= P[Y_\theta \leq L_2] = 1-\beta/2 \\ F_{Y_\theta}(L_1) &= P[Y_\theta \leq L_1] = \beta/2. \end{aligned}$$

The distribution function for Y_θ is F_{Y_θ} and may be used to find the values for the tail points, $y_{\theta,1-\beta/2}$ and $y_{\theta,\beta/2}$ (or L_1 and L_2 depending on your point of view).

The last problem to be faced is the conversion of $L_1 \leq Y_\theta \leq L_2$ into $a \leq \theta \leq b$. The most desireable case occurs when the manipulations required to change $L_1 \leq Y_\theta \leq L_2$ into $a \leq \theta \leq b$ result in a final equation that looks like $P[L_1 \leq Y_\theta \leq L_2] = P[a \leq \theta \leq b]$. In this case $1-\beta = 1-\alpha$ (i.e. $\alpha = \beta$) and the determination of the interval is complete. This general recipe for constructing confidence intervals is applied in the following examples to construct confidence intervals that cover three frequently occurring cases.

1.4.2 The Basic Confidence Intervals.

1.4.2.1 A confidence interval around the mean, known variance, Normal samples. In this case the confidence interval is required around μ and the form to be used will be $Y_\mu = (\overline{X}-\mu)/(\sigma/\sqrt{n})$. Since all of the X_i are Normal, $\overline{X}$ is also normal (see last page). The mean of $\overline{X}$ is μ (see 1.3.2.1) and the standard deviation of $\overline{X}$ is $\sigma/\sqrt{n}$ (see 1.3.2.2). Consequently, Y_μ is standard normal, Z.

[1] The difference between a form and an estimator is that an estimator is a statistic whereas a form may include a sample as well as distributional parameters.

Next note that

$$\begin{aligned} P[L_1 \le Y_\mu \le L_2] &= P[L_1 \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le L_2] \\ &= P[L_1\frac{\sigma}{\sqrt{n}} \le \overline{X}-\mu \le L_2\frac{\sigma}{\sqrt{n}}] \\ &= P[\overline{X}-L_2\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X}-L_1\frac{\sigma}{\sqrt{n}}] \\ &= P[a \le \mu \le b], \end{aligned}$$

with $a = \overline{X} - L_2(\sigma/\sqrt{n})$ and $b = \overline{X} - L_1(\sigma/\sqrt{n})$. The manipulations required to change $L_1 \le Y_\mu \le L_2$ into $a \le \mu \le b$ preserved the equality $P[L_1 \le Y_\theta \le L_2] = P[a \le \theta \le b]$ so $\alpha = \beta$ and $L_1 = z_{1-\alpha/2}$ and $L_2 = z_{\alpha/2}$. Note that in this case the distribution of Z is symmetric about the point 0 so that $L_1 = -L_2$.

Summary:

- X_i is a normal sample, with known variance.
- $Y_\mu = \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ is the form and is standard normal (Z).
- $L_1 = z_{1-\alpha/2} = -L_2$ and $L_2 = z_{\alpha/2}$ (since the distribution of Z is symmetric),
- $\overline{X} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ is a $1-\alpha$ confidence interval about the mean.

Note. In using these summaries to work actual problems, the first bullet point gives the conditions that must be met in order to use the confidence interval stated in the last bullet point. Determine which case you have, calculate the values required by the confidence interval (e.g. $\overline{X}, z_{\alpha/2}$, etc.), and compute the end points of the interval.

1.4.2.2 A confidence interval around the mean, known variance, n large. A classical theorem[1] from probability theory tells us that when n is large, $\overline{X}$ is approximately normal irrespective of the common distribution of the samples:

The Central Limit Theorem. *Let X_i be a sample (a set of iid random variables). Then for n large, $\overline{X}$ is approximately normal with mean μ and variance σ^2/n where μ and σ^2 are the common mean and variance of the samples.*

In the previous section, the samples were assumed to be iid normal. Here, that assumption is replaced by this one: the number of iid samples is large. Now, nowhere in section 6.3.2's discussion was the normality of the samples used except in the assertion that $\overline{X}$ was normal.

[1] The Central Limit Theorem was established in a simplier version by DeMoivre in 1733. Laplace extended DeMoivre's work and was able to state the theorem in the form given here but was unable to prove it. This was accomplished by Liapounoff around 1902.

If the normality of $\overline{X}$ is asserted from the start by appealing to the Central Limit Theorem, the derived confidence interval will be the same as before:

Summary:

- X_i is a sample from any distribution with a large number, n, of samples and known variance.
- $Y_\mu = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ is the form and is standard normal (Z).
- $L_1 = z_{1-\alpha/2} = -L_2$ and $L_2 = z_{\alpha/2}$ (since the distribution of Z is symmetric about 0),
- $\overline{X} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ is a $1-\alpha$ confidence interval about the mean.

1.4.2.3 A confidence interval around the variance, normal samples. In this case the parameter of interest is the variance, σ^2. The form to be used will be

$$Y_{\sigma^2} = \frac{S^2(n-1)}{\sigma^2}.$$

This form works because it is known that the distribution of $S^2(n-1)/\sigma^2$ is $\mathcal{X}^2$ with $n-1$ degrees of freedom (often written $\mathcal{X}^2_{n-1}$) provided that the samples are iid normal. Hence $L_1 = \chi^2_{n-1,1-\beta/2}$ and $L_2 = \chi^2_{n-1,\beta/2}$. In this case also, the transformation of $L_1 \leq Y_{\sigma^2} \leq L_2$ into $a \leq \sigma^2 \leq b$ preserves the probabilistic equality $P[L_1 \leq Y_{\sigma^2} \leq L_2] = P[a \leq \sigma^2 \leq b]$ so $\alpha = \beta$. The values of a and b are easily obtained: $L_1 \leq S^2(n-1)/\sigma^2$ implies $\sigma^2 \leq S^2(n-1)/L_1 = b$ and similarly $S^2(n-1)/\sigma^2 \leq L_2$ implies $a = S^2(n-1)/L_2 \leq \sigma^2$.

Summary:

- X_i is a normal sample,
- $Y_{\sigma^2} = \dfrac{S^2(n-1)}{\sigma^2}$ is the form and is $\mathcal{X}^2$ with $n-1$ degrees of freedom,
- $L_1 = \chi^2_{n-1,1-\alpha/2}$ and $L_2 = \chi^2_{n-1,\alpha/2}$,
- $S^2(n-1)/\chi^2_{n-1,\alpha/2} \leq \sigma^2 \leq S^2(n-1)/\chi^2_{n-1,1-\alpha/2}$ is a $1-\alpha$ confidence interval about the variance.

1.4.2.4 A confidence interval around the mean, unknown variance, normal samples. In this case the parameter of interest is the mean, μ. The form to be used will be

$$Y_\mu = \frac{\overline{X} - \mu}{S/\sqrt{n}}.$$

This form is used because it is known that the distribution of $(\overline{X} - \mu)/(S/\sqrt{n})$ follows Students–t distribution with $n-1$ degrees of freedom (T_{n-1}) provided that the samples are

iid normal. It is known that the T distribution is symmetric about 0 and as n tends to infinity, the distribution tends to the standard normal distribution Z.

The algebraic format of this form is the same as in the case of a confidence interval around the mean with known variance — the σ part in the *known variance* case has been replaced by S here, changing the distribution from Standard Normal (Z) to T_{n-1}. Hence, the only things that change are the tail points.

Summary:

- X_i is a normal sample of unknown variance,
- $Y_\mu = \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ is the form and follows a T distribution with $n-1$ degrees of freedom.
- $L_1 = t_{n-1,1-\alpha/2} = -L_2$ and $L_2 = t_{n-1,\alpha/2}$ (since the distribution of T_{n-1} is symmetric about 0),
- $\overline{X} - t_{n-1,\alpha/2}\dfrac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{n-1,\alpha/2}\dfrac{S}{\sqrt{n}}$ is a $1-\alpha$ confidence interval about the mean when the variance in unknown.

1.4.2.5 A confidence interval for proportions. The parameter of interest in this case is the single trial probability of success, p in a series of Bernoulli trials. Typically in these experiments, a test is repeated n times and a count is kept of the number of successful outcomes of the test. The Binomial random variable, B, counts the number of successes in n Bernoulli trials so the form B/n might serve well as a starting point for creating a confidence interval around p.

The samples used here are the random variables X_i, which take the value 1 if the i-th test was successful but 0 otherwise. Clearly this is a sample (the variables are iid — the distribution in this case is called, strangely enough, a Bernoulli distribution). The density function for each of these Bernoulli random variables is clearly:

$$f(x) = \begin{cases} 1-p & \text{if the test fails (i.e. } X_i = 0) \\ p & \text{if the test succeeds (i.e. } X_i = 1) \end{cases}$$

Note that $B = \sum_{i=1}^{n} X_i$ and taking $\widehat{p} = \dfrac{B}{n}$ yields $\widehat{p} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i = \overline{X}$.

If n is assumed to be large, the confidence interval for the mean with known variance and a large number of samples may be used here because $\widehat{p} = \overline{X}$. Under the large sample size assumption, $\widehat{p}$ is approximately normal with mean equal to the common mean X_i and variance equal to the common variance X_i divided by n. This common mean and common variance of X_i is easily computed:

$$\begin{aligned} \mu &= 0(1-p) + 1p = p \\ \sigma^2 &= (0^2(1-p) + 1^2 p) - (p)^2 = p(1-p) \end{aligned}$$

Summary:

- X_i is a large sample of Bernoulli random variables with $\mu = p$ and $\sigma^2 = p(1-p)$,
- $Y_p = \dfrac{\overline{X} - \mu}{\sigma^2/\sqrt{n}}$ is the form and follows a Z distribution.
- $L_1 = z_{1-\alpha/2}$ and $L_2 = z_{\alpha/2} = -L_1$ (since the distribution of Z is symmetric about 0),
- $\widehat{p} - z_{\alpha/2}\sqrt{\dfrac{p(1-p)}{n}} \leq p \leq \widehat{p} + z_{\alpha/2}\sqrt{\dfrac{p(1-p)}{n}}$ is a $1-\alpha$ confidence interval about the single trial probability of success.

Note: the variance isn't known (it involves p, the quantity being estimated). It is possible to isolate p (in the two inequalities shown in the last summary item) to arrive at an actual confidence interval that surrounds p. However, it has been discovered that only a small error occurs if p is replaced by $\widehat{p}$ in the left and right sides of the confidence interval:

$$\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq p \leq \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

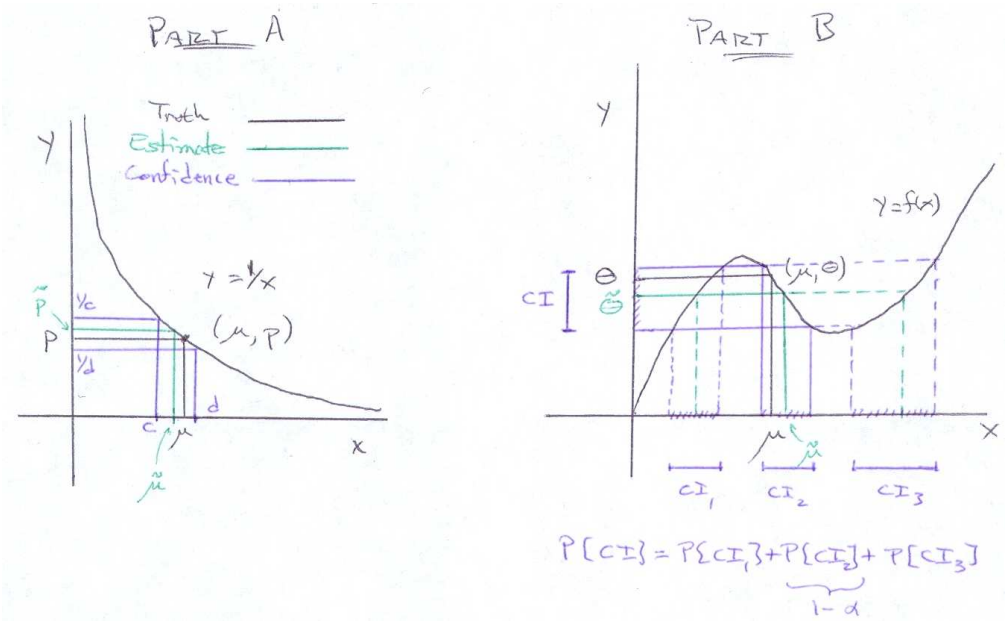
This last expression is commonly used as a confidence interval around p rather than the last summary item above.

1.4.3 Applications – CI around a parameter. Suppose there exists a simple relationship between a parameter of a distribution and the mean that same distribution. As an example, consider the Poisson distribution. To specify a Poisson distribution requires assigning a value to the parameter k. If μ is the mean of a Poisson distribution with parameter k, it is known from the theory that $\mu = k$. The parameter k may be estimated by estimating the parameter μ and it is known that a good estimator for μ is $\bar{X}$. Hence, $\widehat{k} = \widehat{\mu} = \bar{X}$ and $\widetilde{k} = \widetilde{\bar{X}}$. So if $[c\,,\, d]$ is a confidence interval around μ at the $100(1-\alpha)\%$ confidence level, then $[c\,,\, d]$ must be a $100(1-\alpha)\%$ confidence interval around k.

As a second example, suppose X is Geometrically distributed. The parameter required to specify this distribution is p the single trial probability of success. From theory, the mean of X is $\mu = E[X] = 1/p$. Therefore, $p = 1/\mu$. The three basic, generic, $100(1-\alpha)\%$ confidence intervals around the mean all have the same shape: $[c\,,\, d]$ with $c = \bar{X} - w$ and $d = \bar{X} + w$ ($2w = d - c$: the width of the interval). An estimate for μ is $\widetilde{\mu} = \widetilde{\bar{X}}$ so an estimate for p must be $\widetilde{p} = 1/\widetilde{\mu}$. Similarly, if $[c\,,\, d]$ is a $100(1-\alpha)\%$ confidence interval around μ then $[1/d\,,\, 1/c]$ must be a $100(1-\alpha)\%$ confidence interval around the parameter p because:

$$P[1/d \leq p \leq 1/c] = P[d \geq 1/p \geq c] = P[c \leq \mu \leq d] = 1 - \alpha.$$

In this last example, notice that point (μ, p) lies on the curve $y = 1/x$ (see figure below, part A). Once a candidate confidence interval was located for the parameter p, it remainded to verify that this interval was still at the $100(1-\alpha)\%$ level. This may not always be the case



(see figure below, part B). What often happens is that the probability associated with a parameters confidence interval (CI) is split into a sum of several terms ($P[\mathrm{CI}_1], P[\mathrm{CI}_2], P[\mathrm{CI}_3]$ in the figure). One of these probabilities is $1-\alpha$ (the term associated with the confidence interval around μ ... in the figure this is the interval CI_2). The sizes of the other summands are usually very small and only serve to increase the probability (by a tiny amount) that the candidate interval traps the parameter. If however, the sizes of the other summands are not small enough to ignore, the original problem may be reworked with a smaller initial confidence level $(1-\alpha_0)$ and still end with a confidence interval around the parameter at the $100(1-\alpha)\%$ level.

1.5 Hypothesis Testing.

1.5.1 The basic model. A hypothesis test is concerned with establishing the likelihood of a quantitative statement about a parameter θ. Typically the statement will take one of these forms:

$$\theta > \theta_0 \quad \theta \neq \theta_0 \quad \theta < \theta_0$$

where θ_0 is some fixed (and specified) constant.

The Hypothesis. Suppose the issue at hand, the one for which substantiation is desired, is $\theta > \theta_0$. This statement is called the *research hypothesis* (also called the *alternative hypothesis*) and is usually designated by the shorthand H_1. The negation of the research hypothesis ($H_1 : \theta > \theta_0$) is $\theta \leq \theta_0$ and is called the *null hypothesis*. The null hypothesis is usually designated by H_0. Together the pair is usually specified by writing:

$$\begin{aligned} H_1 &: \theta > \theta_0 \\ H_0 &: \theta \leq \theta_0. \end{aligned}$$

The value θ_0 is called the *null value* and is always included with the null hypothesis. Summarizing the three cases, the *alternative/research* hypothesis look like this:

$$\begin{array}{lccc} H_1 : & \theta > \theta_0 & \theta \neq \theta_0 & \theta < \theta_0 \\ H_0 : & \theta \leq \theta_0 & \theta = \theta_0 & \theta \geq \theta_0 \end{array}$$

The Test. Following the designation of the hypothesis, a test is constructed based on an estimator, $\widehat{\theta}$, and a form, Y_θ. If the parameter $\theta = \mu$ is the mean of some random variable the estimator used is often the sample mean ($\widehat{\mu} = \overline{X}$) and the form selected is one of

$$Y_\mu = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \qquad \text{or} \qquad Y_\mu = \frac{\overline{X} - \mu}{S/\sqrt{n}}.$$

Likewise, if $\theta = \sigma^2$ the sample variance S^2 is often taken as the estimator with a form choice of $Y_{\sigma^2} = S^2(n-1)/\sigma^2$.

Decision Criteria. The test protocol itself usually consists of the collection and analysis of sample data as follows. A region $\mathcal{R}$ (an interval of numbers) is selected that will act as an artiber for acceptance or rejection of the null hypothesis. To wit, if the test estimator ($\widehat{\theta}$) produces an estimate ($\widetilde{\theta}$) that falls into the region $\mathcal{R}$, the null hypothesis is automatically rejected. The region $\mathcal{R}$ is called the *critical region* or the *rejection region* for the test. How the region $\mathcal{R}$ is determined is explained below.

Test Results. After the test is run, the experiment will be in one of four states:

- $\widetilde{\theta} \in \mathcal{R}$ and H_0 is false,
- $\widetilde{\theta} \in \mathcal{R}$ and H_0 is true (Type I error),
- $\widetilde{\theta} \notin \mathcal{R}$ and H_0 is false (Type II error),
- $\widetilde{\theta} \notin \mathcal{R}$ and H_0 is true,

The first and last of these states are good because the experiment and reality match. The middle two states are not good since the value or $\widetilde{\theta}$ will cause the null hypothesis to be rejected/accepted in error. These errors are called Type I and Type II errors respectively as shown:

	$\widetilde{\theta} \in \mathcal{R}$	$\widetilde{\theta} \notin \mathcal{R}$
H_0 is true	**Type I**	**Good**
H_0 is false	**Good**	**Type II**

Two probabilities associated with a hypothesis test are required to be known. There are α (called the significance of the test), defined as the probability of making a Type I error, and β, defined as the probability of making a Type II error. In symbols:

$$\alpha = P[\text{Type I error}] = P[\widehat{\theta} \in \mathcal{R} | H_0 \text{ is true}]$$
$$\beta = P[\text{Type II error}] = P[\widehat{\theta} \notin \mathcal{R} | H_0 \text{ is false}]$$

Note that $\widehat{\theta}$ is used in these definitions rather than $\widetilde{\theta}$ because these probabilities are developed before any test is conducted and must stand irrespective of the value actually assumed by $\widehat{\theta}$ after the samples are gathered.

Determination of α. Consider a hypothesis test of the form:

$$H_0 \ : \ \theta \leq \theta_0$$
$$H_1 \ : \ \theta > \theta_0.$$

In support of this hypothesis test, a form (Y_θ) is desired that has the property that as θ moves from ∞ down to $-\infty$ the right tails of Y_θ decrease monotonically in area from a maximum size of 1 down to 0. Imagine the one particular density function associated with Y_θ corresponding to $\theta = \theta_0$ (for this discussion, call this the null value density). All of the density functions associated with the null hypothesis ($\theta \leq \theta_0$) have the bulk of their area to the left of the null value — very little of the density is in the right tails. Just the opposite holds for the research hypothesis theta values — most of their density is to the right of the null value density with little on the left.

In this case, what conclusions can be drawn if the test is run and the test estimator, $\widehat{\theta}$, produces a value that greatly exceeds θ_0? Either (1) the null hypothesis is true and the large positive value was produced by random chance, or (2) the research hypothesis is true

and the large positive value is what would be normally expected. Consequently, a critical region of the form $[C, \infty)$ may be a good choice for $\mathcal{R}$ — the value of C chosen large enough to mitigate case (1) to our satisfaction.

With this done, the probability of a Type I error may be computed:

$$\begin{aligned}\alpha &= P[\widehat{\theta} \in \mathcal{R} | H_0 \text{ is true}] \\ &= P[\widehat{\theta} \geq C \,|\, \theta = \theta_0] \\ &= P[Y_\theta = \mathcal{F}(\widehat{\theta}) \geq \mathcal{F}(C) \,|\, \theta = \theta_0],\end{aligned}$$

where $\mathcal{F}(\theta)$ represents a function that transforms $\widehat{\theta}$ into Y_θ. For example, if $\theta = \mu$ with choice of estimator being $\widehat{\theta} = \overline{X}$ and choice of form being

$$\frac{\overline{X} - \mu}{S/\sqrt{n}},$$

then

$$\mathcal{F}(\theta) = \frac{\theta - \mu}{S/\sqrt{n}}.$$

Note that, in the definiiton of α, the conditional $\theta \leq \theta_0$ has been replaced by $\theta = \theta_0$ because this condition represents the worst case value of α given how the density function of Y_θ behaves as θ decreases. Since the distribution for Y_θ is known the value for α is easy to compute from the value of C or vice-versa.

In a hypothesis test, the significance level of the test (α, the probability of a Type I error) is usually set first and then the boundary value C of the region $\mathcal{R}$ is determined. Typically, setting α too low will result in a very large value for C. Acceptance of the research hypothesis will then occur only in cases when the true value for θ is quite a bit larger than θ_0.

If the original hypothesis is reversed

$$\begin{aligned}H_1 &: \theta < \theta_0 \\ H_0 &: \theta \geq \theta_0.\end{aligned}$$

the same analysis may be used. The critical region will now look like $(-\infty, C\,]$. And if the research hypothesis is $\theta \neq \theta_0$, the critical region will have the shape $(-\infty, C\,] \cup [D, \infty)$ where C and D are usually placed symmetrically on either side of θ_0; that is, $C = \theta_0 - \Delta$ and $D = \theta_0 + \Delta$.

Determination of β. The probability of a Type II error is somewhat more difficult to determine.

$$\beta = P[\widehat{\theta} \notin \mathcal{R} \,|\, H_0 \text{ is false}]$$

The analysis performed in the determination of α set a critical region for the problem. In terms of the same null and alternative hypotheses as before:

$$\begin{aligned}H_1 &: \theta > \theta_0 \\ H_0 &: \theta \leq \theta_0,\end{aligned}$$

the parameter β is now $\beta = P[Y_\theta < \mathcal{F}(C) \,|\, \theta > \theta_0]$. There is no single value of θ, satisfying $\theta > \theta_0$, that represents the worst case value for β. While θ_0 is in some sense a limiting case for the value of θ, substituting $\theta = \theta_0$ here will yield $\beta = 1 - \alpha$ (almost certainly a large value for β). But notice that as θ moves farther to the right of θ_0 the value of β shrinks.

For some value of θ (call it θ_1) the value of β will reach an acceptably small threshold. This threshold value shows how well this test can discriminate between the true value of the parameter θ and the hypothesized value θ_0. If the true value of θ lies to the right of θ_1 the test produces results with acceptably small values for both α and β. If the true value happens to lie between θ_0 and θ_1 Type I error probabilities are still small but Type II error probabilities become large depending on how close the true value θ lies to θ_0. In the region between θ_0 and θ_1 the test's ability to discriminate whether H_0 is true or false is compromised. However, if the form Y_θ is also sensitive to the number of samples, increasing the sample size n may result in better discrimination (the amount of improvement will depend on the statistic).

At the point $\theta = \theta_1$, the corresponding β is used to form the quantity $\rho = 1 - \beta$, called the power of the test. Large values for ρ (that is, small values for β) are usually desired.

A Recap. Here are the steps in conducting a hypothesis test:

- Choose a null and research hypothesis for θ putting the null value θ_0 in with the null hypothesis.
- Pick an estimator for θ and a form associated with θ.
- Choose a value for α, the probability of a Type I error and use it to determine the boundary of the critical region $\mathcal{R}$. Determine a value for θ_1 and compute the power of the test.
- Run the test (this is always done last in a hypothesis test).

The last step is always done on auto-pilot since the key decisions for the test have already been made. The statistic will either fall into the critical region or it will not. On this basis, H_0 is either accepted or rejected.

1.6 Significance Testing. A significance test is a hypothesis test with one difference. A critical (rejection) region is not established before the test begins. Instead, the value taken by the test estimator ($\widetilde{\theta}$) is assumed to lie on the boundary of the critical region. From this assumption, the value of α is computed after the test is over.

Consider the following. A hypothesis test is performed with α set at 0.0001 and after performing the test it is found that the estimate $\widetilde{\theta}$ just missed falling into the critical region (the null hypothesis is accepted). Someone notices that had the value of α been relaxed a bit (say to 0.001), the size of the critical region would change just enough so that now $\widetilde{\theta}$ is in $\mathcal{R}$. This would radically alter the conclusion (H_0 is now rejected) in the aftermath of only a small change in α. Under these circumstances it may be hard to argue that such a small change in the level of significance warrants such a radical change in the test results (accepting the research hypothesis versus rejecting it).

Consequently, when performing a significance test, the key parameter of the test that must be reported is the α value (called the P-value in the parlance of significance testing). The recipients of the test results now must decide whether this P-value makes the the probability of a Type I error unacceeptably large. Following this protocol eliminates the hypothesis test difficulty described in the previous paragraph.

Notice that when conducting a significance test, the design of the test forces the rejection of the null hypothesis because the boundary of the critical region is forced to be the value $\widetilde{\theta}$ and that places the the estimate for θ in the critical (aka rejection) region. Out of this observation three conclusions may be drawn:

- if the P-value is acceptably small, reject the null hypothesis,
- if the P-value is overly large, the accept the null hypothesis,
- if the P-value is neither acceptably small nor overly large, the test in inconclusive (more samples are warrented).

The first of these three statements is clear — $\widetilde{\theta}$ is in $\mathcal{R}$ so reject with a small probability of a Type I error. In the second case, $\widetilde{\theta}$ is still in $\mathcal{R}$ so the null hypothesis is rejected but with a large probability of making aType I error. That is to say, there is a large probability that the null hypothesis is being rejected when it is true. Therefore, the null hypothesis should be accepted. In the third case, the P-value, the probability of a Type I error, is neither large enough nor small enough to make an inference about the truth or falsity of the null hypothesis — the test in inconclusive.

1.7 Commonly Used Statistics. When performing a hypothesis/significance test on the mean one of the best estimators to use is $\overline{X}$. In computing α, transform $\overline{X}$ into a standard normal (Z) form if the variance is known:

$$Y_\mu = Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

If the variance is not known, transform $\overline{X}$ into a T_{n-1} distribution as follows:

$$Y_\mu = T_{n-1} = \frac{\overline{X} - \mu}{S/\sqrt{n}}.$$

If the hypothesis/significance test is on the variance, the S^2 estimator is a popular choice with a Chi-squared form given by

$$Y_{\sigma^2} = \mathcal{X}^2_{n-1} = \frac{S^2(n-1)}{\sigma^2}.$$

2 The Last Page

Proposition 1. *The Sample Mean is an unbiased estimator for the mean.*

Proof: Apply the espectation operator to $\widehat{\mu}$ and incorporate the fact that the X_i are iid with common mean μ:

$$\begin{aligned} E[\widehat{\mu}] &= E[\frac{1}{n}\sum_{i=1}^{n} X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \mu \end{aligned}$$

∎

Proposition 2. *The Sample Variance is an unbiased estimator for the variance.*

Proof: The sample variance is given by:

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\overline{X} + \overline{X}^2\right).$$

Now apply the expectation operation and the fact that the X_i are iid:

$$\begin{aligned} E[\widehat{\sigma}^2] &= \frac{1}{n-1}\sum_{i=1}^{n} E\left[X_i^2 - 2X_i\overline{X} + \overline{X}^2\right] \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(E[X_i^2] - 2E[\overline{X}X_i] + E[\overline{X}^2]\right) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\sigma^2 + \mu^2 - 2E[\overline{X}X_i] + E[\overline{X}^2]\right) \end{aligned}$$

Note that $E[X_i^2] = \sigma^2 + \mu^2$ because X_i comes from the same distribution as X and that $E[\overline{X}] = \mu$ because the previous theorem establishes that $\widehat{\mu} = \overline{X}$ is unbiased.

But what is $E[\overline{X}^2]$ and $E[\overline{X}X_i]$? These are a bit more complicated:

$$\begin{aligned}
E[n^2\overline{X}^2] &= E[\Big(\sum_{i=1}^{n} X_i\Big)^2] \\
&= E[\sum_{i=1}^{n} X_i^2 + \sum_{1\le i\ne j\le n} X_iX_j] \\
&= \sum_{i=1}^{n} E[X_i^2] + \sum_{1\le i\ne j\le n} E[X_iX_j] \\
&= n(\sigma^2+\mu^2) + (n^2-n)\mu^2 \\
&= n\sigma^2 + n^2\mu^2
\end{aligned}$$

So $E[\overline{X}^2] = \sigma^2/n + \mu^2$ and

$$\begin{aligned}
E[n\overline{X}X_i] &= E[X_i\sum_{j=1}^{n} X_j] \\
&= \sum_{j=1}^{n} E[X_iX_j] \\
&= \sum_{1\le j\ne i\le n} \mu^2 + E[X_i^2] \\
&= (n-1)\mu^2 + (\sigma^2+\mu^2) \\
&= \sigma^2 + n\mu^2
\end{aligned}$$

So $E[\overline{X}X_i] = \sigma^2/n + \mu^2$ and

Substituting this into the original expression for $\widehat{\sigma}^2$ produces

$$\begin{aligned}
E[\widehat{\sigma}^2] &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\sigma^2+\mu^2 - 2(\sigma^2/n+\mu^2) + (\sigma^2/n+\mu^2)\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(1-\frac{1}{n})\sigma^2 \\
&= \frac{1}{n-1}(n)(1-\frac{1}{n})\sigma^2 \\
&= \sigma^2
\end{aligned}$$

∎

Proposition 3. *The estimator $\widehat{\mu}_2 - (\widehat{\mu}_1)^2$ is biased for σ^2.*

Proof: It is straightforward to show that $E[\widehat{\mu}_2 - (\widehat{\mu}_1)^2] \neq \sigma^2$:

$$\begin{aligned}
E[\widehat{\sigma}^2] = E[\widehat{\mu}_2 - (\widehat{\mu}_1)^2] &= E[\widehat{\mu}_2] - E[(\widehat{\mu}_1)^2] \\
&= \mu_2 - E[\frac{1}{n^2}\big(\sum_i x_i^2 + \sum_{i\neq j} X_i X_j\big)] \\
&= \mu_2 - \big(\sum_i E[x_i^2] + \sum_{i\neq j} E[X_i X_j]\big)/n^2 \\
&= \mu_2 - \big(\sum_i (\sigma^2 + \mu_1^2) + \sum_{i\neq j} E[X_i]E[X_j]\big)/n^2 \\
&= \mu_2 - \big(\sum_i (\sigma^2 + \mu_1^2) + \sum_{i\neq j} \mu_1^2\big)/n^2 \\
&= \mu_2 - \big(n(\sigma^2 + \mu_1^2) + (n^2 - n)\mu_1^2\big)/n^2 \\
&= (\mu_2 - \mu_1^2) - \sigma^2/n \\
&= \sigma^2(1 - 1/n)
\end{aligned}$$

Note that this establishes that for large n, the bias in the estimator isn't too great. If the estimator was unbiased, the condition $E[\widehat{\mu}_2 - (\widehat{\mu}_1)^2] = \sigma^2$ would hold. What does hold is $E[\widehat{\mu}_2 - (\widehat{\mu}_1)^2] = \sigma^2 - \sigma^2/n$. The difference between being biased and unbiased (the σ^2/n term tends to zero as $n \to \infty$)

∎

Proposition 4 (The sum of normals is normal). *Let X_i be an set of independent normal random variables with mean μ_i and variance σ_i^2. Then the random variable $Y = a\sum_i X_i + b$ is normal for all real constants a and b.*

Proof: Compute the moment generator for the variable Y:

$$\begin{aligned}
m_Y(t) &= E[e^{tY}] \\
&= E[e^{t(a\sum_i X_i + b)}] \\
&= E[e^{atX_1} e^{atX_2} \dots e^{atX_n} e^{bt}] \\
&= E[e^{atX_1}]E[e^{atX_2}] \dots E[e^{atX_n}]E[e^{bt}] \quad \text{all } X_i \text{ are independent} \\
&= m_{X_1}(at) m_{X_2}(at) \dots m_{X_n}(at) e^{bt} \\
&= e^{\mu_1 at + \sigma_1^2 (at)^2/2} e^{\mu_2 at + \sigma_2^2 (at)^2/2} \dots e^{\mu_n at + \sigma_n^2 (at)^2/2} e^{bt} \\
&= e^{(a\mu_1 + a\mu_2 + \dots + a\mu_n + b)t + (a\sigma_1^2 + a\sigma_2^2 + \dots + a\sigma_n^2)(at)^2/2} \\
&= e^{\mu t + \sigma^2 (at)^2/2},
\end{aligned}$$

where $\mu = a\mu_1 + a\mu_2 + \dots + a\mu_n + b$ and $\sigma^2 = a\sigma_1^2 + a\sigma_2^2 + \dots + a\sigma_n^2$.

Hence, the random variable Y is normal with mean μ and variance σ^2.

∎