Estimation; Sampling; The T distribution

I. Estimation

A. In most statistical studies, the population parameters are unknown and must be estimated. Therefore, developing methods for estimating as accurately as possible the values of population parameters is an important part of statistical analysis.

B. <u>Estimators</u> = random variables used to estimate population parameters

C. <u>Estimates</u> = specific values of the population parameters

1. Estimates need not have a single value; instead, the estimate could be a range of values

2. <u>Point estimate</u> = estimate that specifies a single value of the population

3. <u>Interval estimate</u> = estimate that specifies a range of values

D. <u>Properties of a good estimator</u>. Let θ (this is the Greek letter theta) = a population parameter. Let $\hat{\theta}$ = a sample estimate of that parameter. Desirable properties of $\hat{\theta}$ are:

1. <u>Unbiased</u>: Expected value = the true value of the parameter, that is, $E(\hat{\theta}) = \theta$. For example, $E(\overline{X}) = \mu$, $E(s^2) = \sigma^2$.

2. <u>Efficiency</u>: The most efficient estimator among a group of unbiased estimators is the one with the smallest variance. For example, both the sample mean and the sample median are unbiased estimators of the mean of a normally distributed variable. However, $\overline{X}$ has the smallest variance.

3. <u>Sufficiency</u>: An estimator is said to be sufficient if it uses all the information about the population parameter that the sample can provide. The sample median is not sufficient, because it only uses information about the ranking of observations. The sample mean is sufficient.

4. <u>Consistency</u>. An estimator is said to be consistent if it yields estimates that converge in probability to the population parameter being estimated as N becomes larger. That is, as N tends to infinity, $E(\hat{\theta}) = \theta$, $V(\hat{\theta}) = 0$. For example, as N tends to infinity, $V(\overline{X}) = \sigma^2/N = 0$.

II. Sampling

A. Usually, we do not have data on the entire population. Hence, we must rely on observing a subset (or sample) of the population. Population parameters are estimated using the sample.

B. The sample may not necessarily represent the population though.

1. Nonsampling errors - Can study the wrong population. For example, we might exclude those without phones; or, we might rely on self-selection into the sample. For now, we will assume nonsampling errors are not present.

2. Sampling errors - chance alone can cause the sample estimates to differ from the population parameters.

C. There are many different types of sampling schemes. Different types of samples, and the kinds of issues you have to be concerned with when drawing a sample, are discussed in much greater detail in the Research Methods class. For now, we will primarily concern ourselves with a specific type of sampling scheme, simple random sampling. With simple random sampling, each item in the population has an equal and identical chance of being included in the sample.

III. Sample statistics.

A. Suppose we are planning to take a sample of N independent observations in order to determine the characteristics of some random variable X. Then, X_1, X_2,..., X_N represent the observations in this sample. The probability distribution of X_1 through X_N is identical to the distribution of the population random variable X. We want to use the sample values of X_1 - X_N to determine the characteristics of X. Sample statistics are used as estimates of population parameters. The population mean and variance are among the parameters we want to estimate using a sample.

B. The sample mean is defined as:

$$\overline{X} = \frac{1}{N} \Sigma X_i = \hat{\mu}$$

C. The sample variance is

$$s^2 = \frac{1}{N-1} \Sigma (X_i - \overline{X})^2 = \frac{1}{N-1} (\Sigma X_i^2 - N \overline{X}^2) = \hat{\sigma}^2$$

D. Comments on formulas and notation.

1. Notation used differs greatly from source to source and author to author. It is therefore good to become comfortable with a variety of notations

2. Note the use of the ^ in the notation for both the sample mean and variance. It is very common to use a ^ over a population parameter to represent the corresponding sample estimate.

3. In the sample variance, note that the denominator uses N - 1 rather than N. It can be shown that using N - 1 rather than N produces an unbiased estimate of the sample

variance; that is, $E(s^2) = \sigma^2$. Of course, when N is large, it doesn't make much difference whether you use N or N - 1.

IV. Sampling distributions.

A. Note that different samples could yield different values for $\bar{X}$. For example, if I took a random sample of 5 exam scores, I might get a mean of 94; if I took a different random sample of 5 cases, I might get a mean of 92; and so on. That is, $\bar{X}$ is itself a random variable, which has its own mean and variance.

If we take all possible samples of size N and determine for each sample, the resulting distribution is the probability distribution for $\bar{X}$. The probability distribution of $\bar{X}$ is called a sampling distribution.

B. What is the mean and variance of the sampling distribution for $\bar{X}$? That is, what is the mean and variance of $\bar{X}$?

1. Mean of $\bar{X}$:

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \ldots + X_N}{N}\right) = \frac{1}{N}(\mu_X + \mu_X + \ldots + \mu_X) = \frac{1}{N} * N\mu = \mu = \mu_{\bar{X}} = \mu_x$$

2. Variance of $\bar{X}$:

$$V(\bar{X}) = V\left(\frac{X_1 + X_2 + \ldots + X_N}{N}\right) = \frac{1}{N^2}(\sigma^2_X + \sigma^2_X + \ldots + \sigma^2_X) = \frac{N\sigma^2_X}{N^2} = \frac{\sigma^2_X}{N} = \sigma^2_x$$

3. Standard deviation of $\bar{X}$:

$$SD(\bar{X}) = \frac{\sigma_X}{\sqrt{N}} = \sqrt{\frac{\sigma^2_X}{N}} = \sigma_{\bar{X}}$$

The standard deviation of $\bar{X}$ is referred to as the true standard error of the mean.

C. Question: We know the mean and variance of $\bar{X}$ - but what is the shape?

1. If $X \sim N(\mu, \sigma^2)$, then

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{N}\right), and$$

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{N}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} \sim N(0,1)$$

2. What if X is <u>not</u> distributed normally? According to the Central Limit Theorem - regardless of the shape of the parent population (as long as it has a finite mean μ and variance σ^2) the distribution of $\overline{X}$ will approach $N(\mu, \sigma^2/N)$ as the sample size N approaches infinity. That is, as N ---> ∞, $\overline{X} \sim N(\mu, \sigma^2/N)$.

In practice, 30 is usually a pretty good approximation of infinity!

Example: How $\overline{X}$ comes to be normally distributed as N gets larger and larger. As stated above, according to the Central Limit Theorem - regardless of the shape of the parent population (as long as it has a finite mean μ and variance σ^2) the distribution of $\overline{X}$ will approach $N(\mu, \sigma^2/N)$ as the sample size N approaches infinity. That is, as N ---> ∞, $\overline{X} \sim N(\mu, \sigma^2/N)$. We will now give an example of this, showing how the sampling distribution of $\overline{X}$ for the number of pips showing on a die changes as N changes from 1, 2, 3, 10, and 1000. Note how, as N gets bigger and bigger, the distribution of $\overline{X}$ gets more and more normal-like. Also, the true SE of the mean gets smaller and smaller, causing $\overline{X}$ to vary less and less from the true mean of 3.5.

N = 1. When N = 1, $\overline{X}$ and X have the same distribution. This distribution is clearly NOT normal. Rather, it has a *Uniform* distribution – each possible value is equally likely. For example, you are just as likely to get a 1 as you are a 4, even though 4 is much closer to the mean (3.5) than 1 is. By way of contrast, with a normally distributed variable, values close to the mean are more likely than values farther away from the mean.

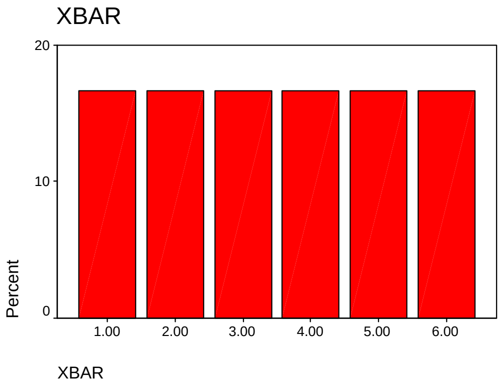
Statistics

XBAR

N	Valid	6
	Missing	0
Mean		3.500
True SE		1.708
Percentiles	16.667	1.167
	33.333	2.333
	50	3.500
	66.667	4.667
	83.333	5.833

XBAR

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	1	16.7	16.7	16.7
	2.00	1	16.7	16.7	33.3
	3.00	1	16.7	16.7	50.0
	4.00	1	16.7	16.7	66.7
	5.00	1	16.7	16.7	83.3
	6.00	1	16.7	16.7	100.0
	Total	6	100.0	100.0	



Cases weighted by WGT

N = 2. Notice how you are already getting big changes – a mean of 1 is now much less likely than a mean of 3.5 or 4. The distribution is not quite normal yet, but it certainly is not uniform anymore. As you see from the percentiles, the middle two-thirds of the distribution runs from 2.083 to 4.917, i.e. if we repeatedly drew samples of size 2, 2/3 of the times the sample mean would be between 2.083 and 4.917 (which in practice is 2.5 to 4.5 inclusive).

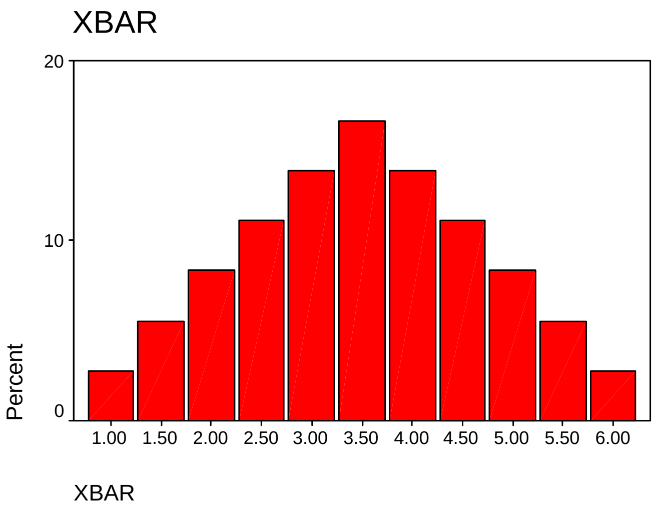
Statistics

XBAR

N	Valid	36
	Missing	0
Mean		3.500
True SE		1.208
Percentiles	16.667	2.083
	33.333	3.000
	50	3.500
	66.667	4.000
	83.333	4.917

XBAR

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	1	2.8	2.8	2.8
	1.50	2	5.6	5.6	8.3
	2.00	3	8.3	8.3	16.7
	2.50	4	11.1	11.1	27.8
	3.00	5	13.9	13.9	41.7
	3.50	6	16.7	16.7	58.3
	4.00	5	13.9	13.9	72.2
	4.50	4	11.1	11.1	83.3
	5.00	3	8.3	8.3	91.7
	5.50	2	5.6	5.6	97.2
	6.00	1	2.8	2.8	100.0
	Total	36	100.0	100.0	



Cases weighted by WGT

N = 3. The distribution continues to get more and more normal like, and values further from the mean continue to get less and less likely. The middle 2/3 now covers a smaller range, 2.67 through 4.33.

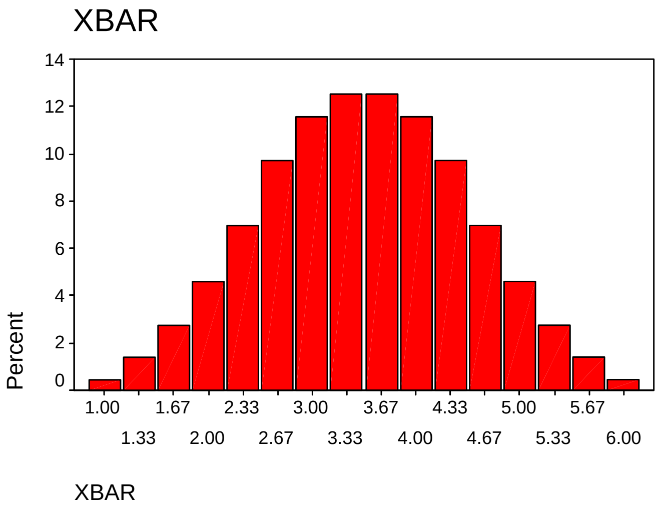
Statistics

XBAR

N	Valid	216
	Missing	0
Mean		3.500
True SE		.986
Percentiles	16.667	2.667
	33.333	3.000
	50	3.500
	66.667	4.000
	83.333	4.333

XBAR

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	1	.5	.5	.5
	1.33	3	1.4	1.4	1.9
	1.67	6	2.8	2.8	4.6
	2.00	10	4.6	4.6	9.3
	2.33	15	6.9	6.9	16.2
	2.67	21	9.7	9.7	25.9
	3.00	25	11.6	11.6	37.5
	3.33	27	12.5	12.5	50.0
	3.67	27	12.5	12.5	62.5
	4.00	25	11.6	11.6	74.1
	4.33	21	9.7	9.7	83.8
	4.67	15	6.9	6.9	90.7
	5.00	10	4.6	4.6	95.4
	5.33	6	2.8	2.8	98.1
	5.67	3	1.4	1.4	99.5
	6.00	1	.5	.5	100.0
	Total	216	100.0	100.0	



Cases weighted by WGT

N = 10. This is very close to being a normal distribution. When N = 1, there is a 1/6th chance that the sample mean would be 1; when N =10, there is less than 1 chance in 60 million of getting a sample mean that small. Now the middle two-thirds runs from 3 to 4. If you rolled a die 10 times, there is only a very small chance (about 0.3%) that you would get a mean of less than 2 or greater than 5.

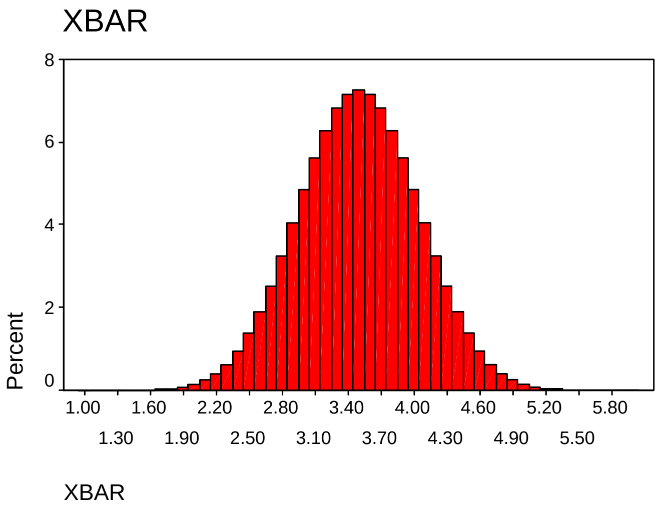
Statistics

XBAR

N	Valid	60466176
	Missing	0
Mean		3.500
True SE		.540
Percentiles	16.667	3.000
	33.333	3.300
	50	3.500
	66.667	3.700
	83.333	4.000

XBAR

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	1	.0	.0	.0
	1.10	10	.0	.0	.0
	1.20	55	.0	.0	.0
	1.30	220	.0	.0	.0
	1.40	715	.0	.0	.0
	1.50	2,002	.0	.0	.0
	1.60	4,995	.0	.0	.0
	1.70	11,340	.0	.0	.0
	1.80	23,760	.0	.0	.1
	1.90	46,420	.1	.1	.1
	2.00	85,228	.1	.1	.3
	2.10	147,940	.2	.2	.5
	2.20	243,925	.4	.4	.9
	2.30	383,470	.6	.6	1.6
	2.40	576,565	1.0	1.0	2.5
	2.50	831,204	1.4	1.4	3.9
	2.60	1,151,370	1.9	1.9	5.8
	2.70	1,535,040	2.5	2.5	8.3
	2.80	1,972,630	3.3	3.3	11.6
	2.90	2,446,300	4.0	4.0	15.7
	3.00	2,930,455	4.8	4.8	20.5
	3.10	3,393,610	5.6	5.6	26.1
	3.20	3,801,535	6.3	6.3	32.4
	3.30	4,121,260	6.8	6.8	39.2
	3.40	4,325,310	7.2	7.2	46.4
	3.50	4,395,456	7.3	7.3	53.6
	3.60	4,325,310	7.2	7.2	60.8
	3.70	4,121,260	6.8	6.8	67.6
	3.80	3,801,535	6.3	6.3	73.9
	3.90	3,393,610	5.6	5.6	79.5
	4.00	2,930,455	4.8	4.8	84.3
	4.10	2,446,300	4.0	4.0	88.4
	4.20	1,972,630	3.3	3.3	91.7
	4.30	1,535,040	2.5	2.5	94.2
	4.40	1,151,370	1.9	1.9	96.1
	4.50	831,204	1.4	1.4	97.5
	4.60	576,565	1.0	1.0	98.4
	4.70	383,470	.6	.6	99.1
	4.80	243,925	.4	.4	99.5
	4.90	147,940	.2	.2	99.7
	5.00	85,228	.1	.1	99.9
	5.10	46,420	.1	.1	99.9
	5.20	23,760	.0	.0	100.0
	5.30	11,340	.0	.0	100.0
	5.40	4,995	.0	.0	100.0
	5.50	2,002	.0	.0	100.0
	5.60	715	.0	.0	100.0
	5.70	220	.0	.0	100.0
	5.80	55	.0	.0	100.0
	5.90	10	.0	.0	100.0
	6.00	1	.0	.0	100.0
	Total	60,466,176	100.0	100.0	



Cases weighted by WGT

N = 1000. I cannot generate enough data to illustrate this! (When N = 10, there are already more than 60 million possible combinations.) But if I could, the mean would be 3.5, and the true standard error would be .054. The middle 2/3 would run from about 3.446 to 3.554. If we took the average of 1000 dice rolls, there is only about a 0.3% chance that the average would not be between 3.34 and 3.66.

V. The T distribution

1. Note that, in order to do a Z-score transformation, σ must be known. In reality, such a situation would be extremely rare. Much more common is the situation where both μ and σ are unknown. What happens if we do not want to (or cannot) assume a value for σ (i.e. σ is unknown)? When σ is unknown, we substitute the sample variance s; and instead of doing a Z transformation which produces a N(0,1) variable, we do a T transformation which produces a variable with a T distribution.

That is, when σ is unknown,

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{N}}} = \frac{\overline{X} - \mu}{\sqrt{\frac{s^2}{N}}}; \quad \frac{\overline{X} - \mu}{\frac{s}{\sqrt{N}}} \sim T_{N-1}$$

Unfortunately, when s is substituted for σ, T is not normally distributed, nor is its variance 1. Since s is itself a random variable, there is more uncertainty about the true value of $\underline{X}$, hence the variance of T is greater than 1.

NOTE: You want to be sure to distinguish between the following:

$$\frac{\sigma}{\sqrt{N}} = True\ standard\ error\ of\ the\ mean,$$

$$\frac{s}{\sqrt{N}} = est\ \sigma_M = Estimated\ standard\ error\ of\ the\ mean$$

2. <u>Characteristics of the T distribution</u>:

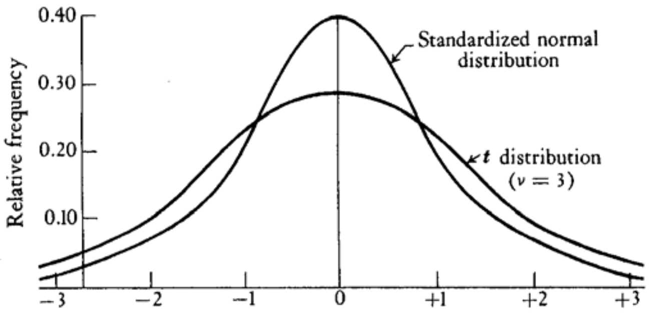
✓ Shape is determined by one parameter, v = N - 1 = degrees of freedom. Why N - 1? If you know s, and the value of N - 1 observations, you can determine the value of the Nth observation. Hence, only N - 1 observations are free to vary.

✓ E(T) = 0.

✓ The T distribution is symmetric.

✓ As N tends to infinity, T ~ N(0, 1). 120 is a pretty good approximation of infinity, and even 30 isn't too bad. (Note: this is fairly logical, since the bigger the sample size, the closer s approximates σ).

✓ The following diagram (from Harnett) shows how a variable with a T distribution and 3 degrees of freedom compares to a variable with a standardized normal distribution. Note that values are much more spread out in the T-distributed variable, i.e. the tails are fatter. As V increases, the T distribution becomes more and more like a standardized normal distribution; after v = 120, there is very little difference, and even after v=30 there is not much difference.



3. The T transformation is appropriate whenever the parent population is normally distributed and σ is unknown. Even if the parent population is not normally distributed, T will often work ok.

4. Tables for the T distribution. T is a little bit harder to work with than Z, because $P(a \leq T \leq b)$ depends upon the degrees of freedom. For example, for v = 120,
$P(-1.98 \leq T \leq 1.98) = .95$, but when v = 10, $P(-2.228 \leq T \leq 2.228) = .95$. Ergo, tables for the T distribution typically list critical values for T at specific levels of significance (i.e. α = .01, α = .05) for different values of v.

See Appendix E, Table III, (or Hays p. 932) for tables for the T distribution.

In Table III, the first column gives values for v. The first row lists significance values for what Hayes labels as Q and 2Q. If you want to do a 1-tailed test, look at the line labeled Q; for a 2-tailed test, use the line labeled 2Q. $Q = P(T_v \geq t) = P(T_v \leq -t)$, and $2Q = 1 - P(-t \leq T_v \leq t)$.

✓ For example, suppose N = 11 (i.e. v = 10) and we want to find t for $P(T_{10} \geq t) = .05$; looking at the row labeled v = 10 and the column labeled Q = 0.05, we see that the critical value is 1.812. (Note that this is larger than the z value of 1.65 that we use when σ is known). Note that this also means that F(1.812) = .95.

✓ Conversely, if we wanted to find t for $P(T_{10} \leq t) = .05$, the appropriate value would be -1.812 (meaning $F(-1.812) = .05$.

✓ Or, suppose $v = 15$, $\alpha = .01$, and we want $P(-a \leq T \leq a) = .99$. We look at the row for $v = 15$ and the column for $2Q = 0.01$, and we see that $a = 2.947$, i.e. $P(-2.947 \leq T_{15} \leq 2.947) = .99$. (By way of contrast, when σ is known, the interval is -2.58 to 2.58).

Other than always having to look up values in the table, working with the T distribution is pretty much the same as working with the N(0, 1) distribution.

EXAMPLES:

1. Find the values of t for the T distribution which satisfy each of the following conditions:

(a) the area between -t and t is 0.90 and v = 25

(b) the area to the left of t is 0.025 and v = 20

Solution.

(a) $P(-t \leq T_{25} \leq t) = .90 \implies 2Q = .10$. Looking at Table III, appendix E, $v = 25$, $2Q = .10$, we see that $t = 1.708$.

(b) $P(T_{20} \leq t) = .025 = P(T_{20} \geq -t) \implies t = -2.086$. (Look at $Q = .025$, $v = 20$.)