Lecture 2 – Inference through estimation

One method of making inference about a population is to estimate the properties of the sample. Questions asked would be: What is the average of the sample? How variable is the sample? If I repeated the sample, would I get a similar result? How sure am I that the population parameter falls within a certain range?

Descriptions of central tendency

There are several measures of the central tendency of a distribution. These are known as *summary statistics*. We will most often use the mean, but the mode and median are used as well.

- Mean – centre of gravity of a distribution. Also called the expected value for a particular distribution.
- Median is where half the data are greater and half less than the value.
- Mode is the highest peak.

Arithmetic mean

There are several "means." In particular we care mainly about the arithmetic mean because of the Normal distribution.

The arithmetic mean is sum of all data divided by the number of observations. The mean is denoted by the “bar” about the variable name.

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ where y_i is the i^{th} observation in the sample of n observations.

Median

The median is the middle measurement in a set of ordered data.

(If there are an even number of data points, the median is usually given as the average of the two middle data points.)

For example:

18 28 28 24 25 36 30 28 14 17 22 34 26 22 20

can be put in order:

14 17 18 20 22 22 24 25 26 28 28 28 30 34 36

Here there are seven data points about 25, and seven below 25, therefore 25 is the median value.

Mode

The mode is the most frequently occurring measurement in a data set or distribution

For example, 28 occurs three time in the above data set, which is more often than any other value, so the mode of this data is 28.

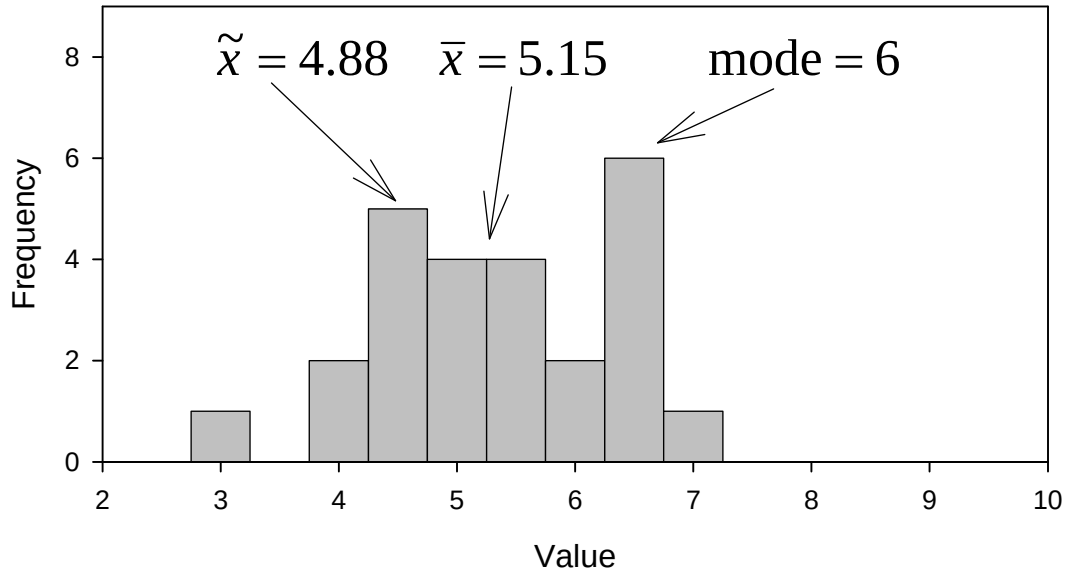

Figure 2.1. Example of the mean, median and mode of a 25 observations drawn from a Normal distribution with a mean of 5 and a standard deviation of 1 or $y \sim N(5,1^2)$.

Descriptions of spread

Range

The simplest measure of the dispersion of a data set is its range. The range is the difference between the maximum and minimum measurement. For example, the range of the data given above in the section on medians is

Range : 36 -14 = 22.

The range is very sensitive to sample size, however - larger samples tend to have a larger range even if they are drawn from the same distribution. This makes the range an unsatisfactory measure of the dispersion of a distribution.

Variance

The most commonly used measure of spread is the variance. Variance is defined as the average squared deviation from the mean, such as

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

where s^2 is the sample variance, and *n* is the number of individuals in the sample.

Standard Deviation

The standard deviation, or SD for short, is just the positive square root of the variance. It expresses exactly the same information as the variance, but re-scaled to be in the same units as the mean.

The sample standard deviation, represented by *s*, is the square root of the sample variance.

(Standard deviations are sometimes more intuitive. Variances have other attractive properties, such as they are additive, that is, the variance of a sum of variables is related to the sum of the variances of those variables.)

Coefficient of Variation

The Coefficient of Variation, or CV for short, is another way of expressing the information in the variance, but standardized to the mean value of the variable. The CV is usually expressed as a percent. Note that this standardization leaves no units on the CV; it is dimensionless.

CV = 100 s / $\bar{x}$

Standard error of the mean

The standard error of the mean, or SE for short, is a measure of the repeatability of an estimate.

Imagine that a sample of the same size as the real was hypothetically taken from the same population an infinite number of times. Each of these imaginary samples would give a different estimate of the parameter in question. The standard deviation of these pseudo-estimates is what is called the standard error of the estimate.

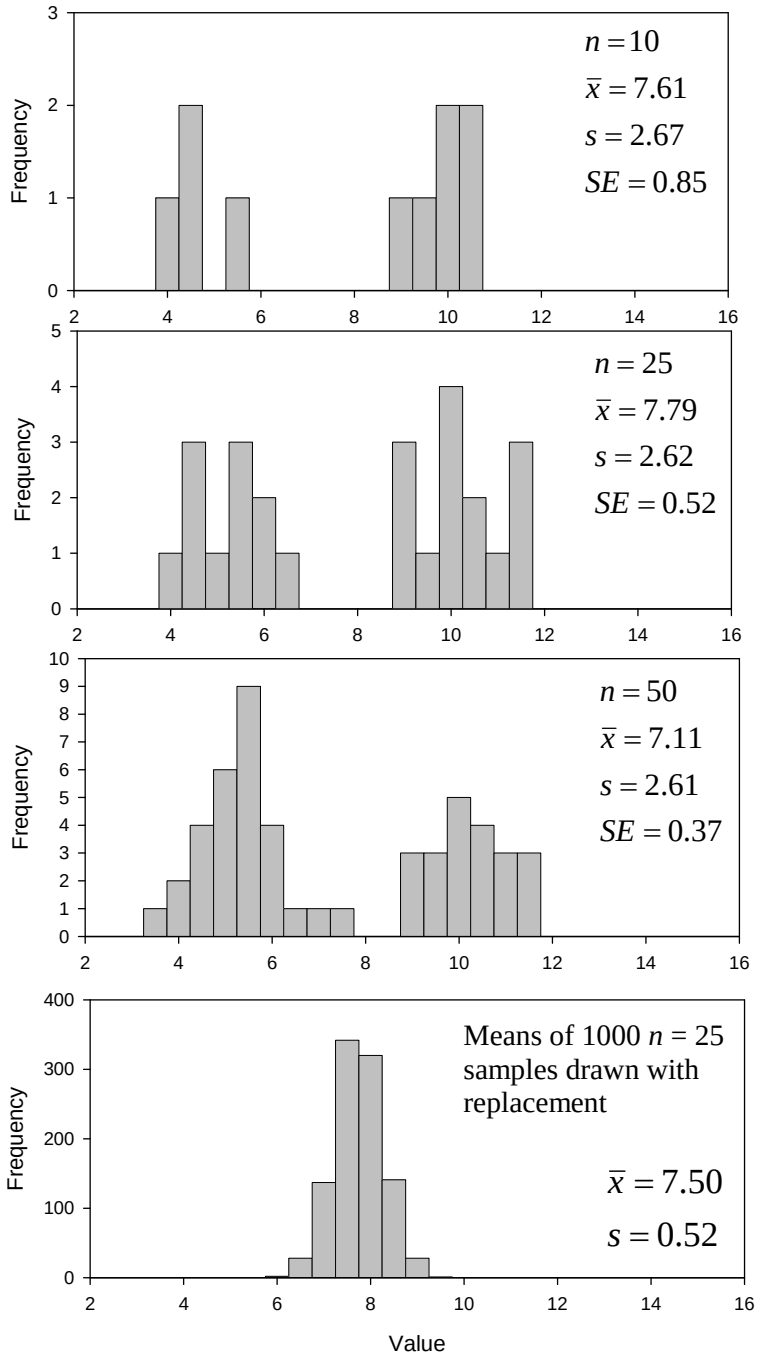

Figure 2.2. These data are generated from an extremely non-normal distribution – a bimodal distribution. Note that the means of the resampled data are Normally distributed and the standard error of the mean of the *n* = 25 sample and the standard deviation of all the means are similar. Note that 1) as the sample size increases how the sample starts to approximate a bimodal distribution, and 2) that as the averages are plotted 1000 times how the standard deviation of these averages is similar to the standard error of the mean in the sample with *n* = 25. The average of the resample sis also 7.5 – the average of the two modes at 5 and 10.

In reality of course we cannot get an infinite number of samples: we only have one, in fact. So we will use our understanding of probability to predict the standard error of any given sample, based on the properties of that sample.

For example, the means of samples of *n* individuals taken from a normal distribution will have a standard deviation equal to $SE = s_{\bar{x}} = \frac{s}{\sqrt{n}}$ so we say that this is the standard error of the mean.

Note that the standard error goes down, and therefore the reliability goes up, if we have larger samples.

The Central Limit Theorem

If we take a large enough samples, the means of the samples will be approximately normally distributed even if the underlying variable being averaged is not normally distributed. This is the Central Limit Theorem. The phrase "large enough" in the previous sentence will vary from distribution to distribution; the further the underlying distribution is from normal, the larger the sample which is required to give a mean which is normally distributed.

The Central Limit Theorem is extremely important to a statistical understanding of the world. It is likely that the reason that many things have a nearly normal distribution is that most of the variables we observe are themselves sums of other underlying variables. Arm length is normally distributed because there are lots of genes and environmental influences that act together to cause an arm to be as it is. Furthermore, as we will see later, many statistical tests assume that means are normally distributed, and the CLT allows us to sometimes use these tests even when the underlying distributions.

Degrees of Freedom

The size of samples used to generate an estimate is often described in terms of the numbers of degrees of freedom in that sample. There as many degrees of freedom in a sample for a particular estimate as there are independent terms used to calculate that estimate. For example, for a mean there are *n* degrees of freedom in a sample of size *n*; for an estimate of variance there are *n*-1 degrees of freedom in that same sample.

This is because the mean is used to generate the estimate of the variance, and therefore only *n*-1 of the data points can vary and still give the same mean.

The distribution of the data vs the distribution of the statistics

It is important to reiterate the importance of not confusing the distribution of the data with the distribution of the statistics. The former is from a data generating function, the *pdf*, that has an unknown mean and variance, while the latter, the sample statistic, is one of an infinite number of possible summary statistics. Fortunately (because of the CLT) we know that given a very,very large number of samples, the sample statistics will be normally distributed. At least one thing is known.

In the case of in data then each datum is Z-distributed

$$Z = \frac{x - \mu}{\sigma}$$

This is also known as a standard normal with a mean of 0 and a variance of 1. This is shown by retransforming the data back as $Z\sigma + \mu = x$ where we scale the statistic with σ and then adding the offset μ.

In the case of the mean **statistic** the statistic is also Z-distributed as

$$Z = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

DO NOT CONFUSE THE TWO

In small sample sizes then the means are t-distributed as

$$T = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where *T* is t-distributed with (*n*-1) degrees of freedom. The t-distribution looks like a Z-distribution but at smaller samples has "fatter" tails. At larger samples (>30) the two are pretty much similar.

The new transformed *T* statistic is a key quantity for use in later inferential statistics.

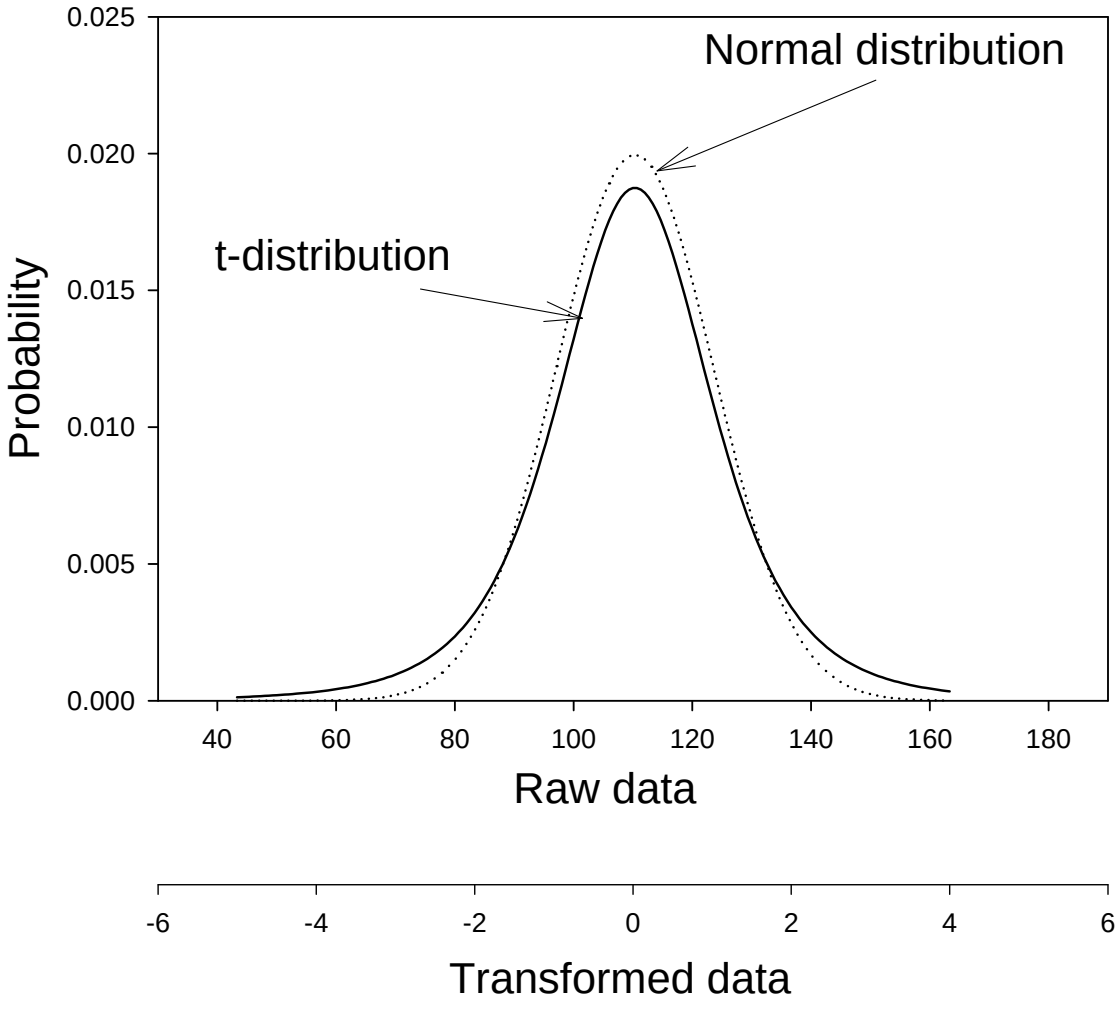

Figure 2.3. If the data were generated from a $N(110,20^2)$ then the data can be transformed easiy to a standard normal (or *z*-distribution) (dotted line) or a *t*-distribution (solid line). Note the "fatter tails" in the *t*-distribution which has been plotted with $n = 5$ degrees of freedom.

Confidence intervals for μ

Once we have information pertaining to central tendency and spread of our sample, we can ask inferential questions about our population. This is through the construction of confidence intervals such that we can state "given our sample, we are 95 percent certain that the population average falls with X and Y".

Confidence intervals are based on the assumption that the sample has been randomly selected from a normal population. It is appropriate for samples of any size and works satisfactorily even if the data are not from a normal population, so long as the departure from normality is not excessive.

Recall that $T = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$ has a t-distribution with (n-1) degrees of freedom. This will serve as a pivotal quantity in forming a confidence interval for μ. From tables we know that we can find values $t_{\alpha/2}$ and $-t_{\alpha/2}$ such that $P\left(-t_{\alpha/2} \le T \le t_{\alpha/2}\right) = 1-\alpha$.

Note: if α is a probability from a distribution, then t_α would be that value on the x-axis from the *t*-distribution where the all the probability to the right of it would sum to α.

For example, from Figure 2.3, the value 1.96 would correspond to an α of 5%. Therefore $z_\alpha = 1.96$ or $P(z_\alpha > 1.96) = 0.05$.

Therefore, if $P\left(-t_{\alpha/2} \le \dfrac{\bar{y} - \mu}{s/\sqrt{n}} \le t_{\alpha/2}\right) = 1-\alpha$ then

$P\left(-t_{\alpha/2}\left(\dfrac{s}{\sqrt{n}}\right) \le \bar{y} - \mu \le t_{\alpha/2}\left(\dfrac{s}{\sqrt{n}}\right)\right) = 1-\alpha$ (multiply by standard error of the mean)

$P\left(-\bar{y} - t_{\alpha/2}\left(\dfrac{s}{\sqrt{n}}\right) \le -\mu \le -\bar{y} + t_{\alpha/2}\left(\dfrac{s}{\sqrt{n}}\right)\right) = 1-\alpha$ (subtract y)

$P\left(\bar{y} - t_{\alpha/2}\left(\dfrac{s}{\sqrt{n}}\right) \le \mu \le \bar{y} + t_{\alpha/2}\left(\dfrac{s}{\sqrt{n}}\right)\right) = 1-\alpha$ (negate and rearrange inequalities)

If values are normally distributed then the confidence in interval for the population mean μ is $\bar{y} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$

The population parameter ranges between the sample mean plus and minus with $t_{\alpha/2,n-1}$ standard errors of that mean.

Confidence interval for differences in two population means $\mu_1 - \mu_2$ is

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2,n_1+n_2-2}\left(\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}\right)\left(\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

The variance is from two samples, so to include possible unequal samples size the pooled we have

$$s^2_{pooled} = \left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right)$$

Inference can be made about μ based on the resultant confidence interval. Does the confidence interval include a particular value? In the case of a question about the differences between two means, say μ_1 and μ_2, then if the confidence interval includes 0 then they are most probably similar.

Example:

Given the following statistics ($\bar{y} = 20$ $s^2 = 25$ $n = 20$) from a sample what is the 95% confidence interval of μ ?

$$\bar{y} - t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right) \le \mu \le \bar{y} + t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

From the tables we have $t_{0.025,19} = 2.093$

$$20 - 2.093\left(\frac{5}{\sqrt{20}}\right) \le \mu \le 20 + 2.093\left(\frac{5}{\sqrt{20}}\right)$$

The 95% confidence interval for μ is $17.66 \le \mu \le 22.34$

Example:

Given the following statistics from two samples what is joint 90% confidence interval of $\mu_1 = \mu_2$? The data are $\bar{y}_1 = 20$ $s_1^2 = 25$ $\bar{y}_1 = 18$ $s_1^2 = 22$ $n_1 = n_2 = 20$.

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2, n_1+n_2-2} \left(\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \right) \left(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

From the tables we have $t_{0.05,38} = 1.645$ (as 38 is greater than 29 on the table we either go for the ∞ option or the z-table. They are equivalent)

$$(20-18) \pm 1.645 \left(\sqrt{\frac{(19)25 + (19)22}{38}} \right) \left(\sqrt{\frac{1}{20} + \frac{1}{20}} \right)$$

The 90% confidence interval for $\mu_1 - \mu_2$ is $-0.52 \le \mu_1 - \mu_2 \le 2.52$

Confidence interval inference

These confidence intervals allow us to make inference statements about μ or $\mu_1 - \mu_2$.

In the former case, we can be 95% certain that the population mean is not 25 as it falls outside the 95% confidence interval.

We can also be 90% sure that the difference between the two population means is zero because zero falls within the 90% confidence interval.