Set 22 and 23

A **point estimate** of a parameter is a single number that serves as an estimate for the true value of the parameter. The estimate comes from a statistic (which we might also call an **estimator**), and is then calculated from the sample data.

Example: We might not know the population mean μ (also called the **true mean**) of some population. We take a sample of n observations, and then find the value of the sample mean $\overline{X}$, which is our estimator. The value we find, $\overline{x}$, is our point estimate for the true mean μ.

Properties of a good estimator:

- The estimator should be **unbiased**; this means that the estimator does not tend to overestimate, and does not tend to underestimate.

  If $\widehat{\theta}$ is an unbiased estimator for some parameter θ, then:

  $$E(\widehat{\theta}) = \theta$$

  In other words: the long-run average value of the estimate will be the parameter which we are estimating.

- The estimator should be **consistent**; this means the value of the estimate will approach the true value of the parameter as $n \to \infty$, where n is the sample size.

Example: Suppose we have a population with an unknown true mean μ. We select n members of the population as our sample, and wish to use $\overline{X}$, the sample mean, as the estimator which will give us the point estimate of our true mean,

Is $\overline{X}$ unbiased for μ? Is $\overline{X}$ a consistent estimator?

Example: Suppose a population has unknown mean μ and variance σ^2. We take n observations, $X_1, \ldots, X_n$ from this population.

We can show that $S^2 = \dfrac{\sum \left(X_i - \overline{X}\right)^2}{n-1}$, the sample variance, is an unbiased estimator for σ^2 (i.e. we can show $E(S^2) = \sigma^2$).

If we have a choice between estimators, we would prefer the estimator with the greater **efficiency**. The less variability an estimator has in estimating the parameter, the more efficient it is.

If two estimators are both unbiased, then the estimator with the smaller variance is the more efficient estimator.

Example: Suppose we have a normal population, with unknown mean μ and variance σ^2. We take two observations X_1 and X_2, selected independently from this population.

One choice of an estimator for μ is $\overline{X} = \dfrac{X_1 + X_2}{2}$.

Another choice of estimator for μ is $Y = 2X_1 - X_2$.

Show that both estimators are unbiased for μ. Decide which estimator is more efficient.

An **interval estimate** of a parameter is an interval of real numbers, where each number in the interval is an estimate for the true value of the parameter.

A **pivotal quantity** is a function of observations with a distribution that does not depend on the value of any unknown parameters.

Example: Suppose we have a random sample of n, either from population with mean μ and standard deviation σ. Also, suppose the population is normal, or that n is large (or both).

Then $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ is a pivotal quantity.

By the Central Limit Theorem, we know that $\overline{X}$ is normal with mean μ and standard deviation $\sigma/\sqrt{n}$. So, the expression above will have standard normal distribution, regardless of what the values of μ, σ are.

We can use these pivotal quantities to construct a **random interval** for one of the parameters. A random interval is an interval of real numbers whose endpoints are random variables.

Example: Suppose we have a random sample of n observations from some normal population. We can find that

$$P\left(-1.96 \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

This can be rewritten as:

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

We have constructed the random interval $\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}\ ,\ \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$. This interval has the property that over many repetitions of the experiment, 95% of the time the true value of μ will lie within the interval.

Comments:

- There are many other choices for a starting upper and lower bound other than ± 1.96 that we could have chosen.

  We could have used the fact that $P(-\infty \leq Z \leq 1.645) = 0.95$, or that $(-1.645 \leq Z \leq \infty) = 0.95$.

  If we had done so, we would have ended up with a **one-sided random interval**, instead of the **two-sided random interval** that we had created.

- My choice of a **coverage probability** of 95% was arbitrary. If I decide I want more of the intervals I create to contain μ, I can start with a higher coverage probability. There are costs to this, which we will discuss later.

Suppose we know the value of σ, make our n observations, and compute the value of $\overline{x}$.

We would call the following *fixed* interval a **95% confidence interval** for μ.

$$\left(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}\ ,\ \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

The lower end of this interval is the **lower confidence limit** and the upper end is the **upper confidence limit**.

In general, the following are the upper and lower limits of a $100(1-\alpha)\%$ CI for μ:

$$\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Recall that $z_{\alpha/2}$ would be the value such that that area to the *right* of it under the standard normal curve is $\alpha/2$.

We call $z_{\alpha/2}$ the **critical value** for the confidence interval.

Our confidence interval for μ consists of

- $\overline{x}$, the point estimate for μ.
- a critical value from Z
- the standard error of $\overline{x}$.

In several different scenarios, the pivotal quantity from which we derive our confidence interval has standard normal distribution (or another symmetric distribution). In all such cases, the confidence interval is of the form:

$$\text{estimate} \pm (\text{critical value})(\text{standard error})$$

Common critical values:

% Confidence	Critical Value
90	$z_{\alpha/2} = z_{0.05} = 1.645$
95	$z_{\alpha/2} = z_{0.025} = 1.96$
99	$z_{\alpha/2} = z_{0.005} = 2.575$

Note: While these confidence levels are common, you should also be able to find the critical value for *any* confidence level.

Example: Suppose that for 25 people, the average height $\overline{x} = 1.7$ m. Suppose it is known that the standard deviation of people's heights is 0.08 m, and that people's heights are normally distributed. Find a 95% confidence interval for μ, the population's mean height.

Interpretation: We should **not** interpret this as meaning that there is a 95% chance that the true mean height is between 1.66864 and 1.73136 meters.

Instead, we should interpret the idea of "95% confidence" as meaning that if we were to repeat this experiment many times, about 95% of the intervals we would construct would contain the true value of μ.

The level of confidence we have is measuring our confidence in our **procedure**, not in the particular interval we've created.

We should note that the d, the distance from the center to the lower or upper confidence limit (also called the **margin of error**) depends on the critical value and the standard error.

$$d = (\text{critical value})(\text{standard error})$$

$$d = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Note: We can also talk about the **width** of a confidence interval, which is the distance between the upper and lower confidence limits. That is, $2d$ is the width of a CI.

If the standard deviation is known (or assumed), and the critical value is fixed in advance, we can determine the sample size needed to give us a particular margin of error, d.

Example: Suppose the lifetime of certain type of lightbulb is normally distributed and has a standard deviation of $\sigma = 200$ hours. How many samples do we need to be create a 95% confidence interval for μ, the mean lifespan, with a margin of error of 10 hours?

Example: How many observations do we need to create a 95% confidence interval for μ with a width of 40 hours?

In most real-life cases, the true values of σ and μ are both unknown. However, if our sample size is quite large, then s, the sample standard deviation, should be a good estimate for σ.

For a large-sample size scenario ($n \geq 40$), the following is an approximate $100(1-\alpha)\%$ confidence interval:

$$\left(\overline{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

We call $s/\sqrt{n}$ the **estimated standard error** of the sample mean, since it is an estimate of $\sigma/\sqrt{n}$, the standard error of the sample mean.

Example: At a particular location, fifty daily measurements of wind speed (in m/s) are made. It is found that $\overline{x} = 15.9\ m/s$ and $s = 7.7\ m/s$.

Find a 98% confidence interval for μ, the average daily wind speed. Assume that the measurements constitute a random sample from the population of all wind speed measurements.

Suppose $X_1, X_2, \ldots, X_n$ are a random sample from a normal population. Also, suppose that the sample size is not large.

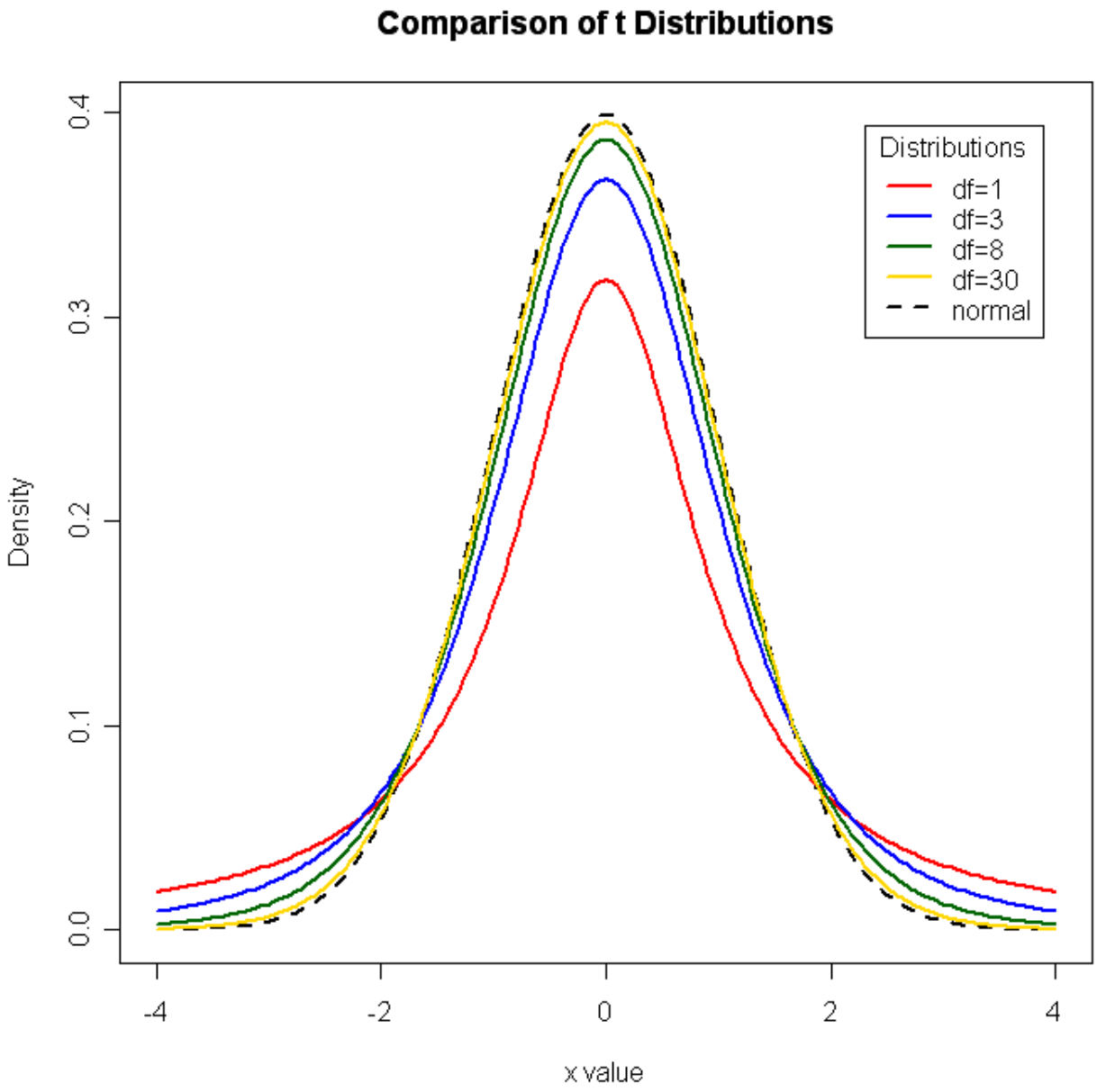
The following pivotal quantity (where $\overline{X}$ is the sample mean, and S is the sample standard deviation) will have t-distribution, with $n-1$ degrees of freedom.

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

Properties of the t distribution:

1. The t distribution is continuous, and defined on $(-\infty, \infty)$.
2. The t distribution is symmetric, bell-shaped, and centered at zero.
3. The number of degrees of freedom affect the shape of the distribution; as the number of degrees of freedom increases, the distribution becomes more peaked, and the tails become thinner.
4. When the number of degrees of freedom is large (30 or more), the t-distribution is approximately a standard normal distribution.

Notation: To denote a t-distribution with k degrees of freedom, we write t_k.


Comparison of t Distributions
Distributions
df=1
df=3
df=8
df=30
normal
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x value

Confidence Interval for the Population Mean: The $100(1-\alpha)\%$ confidence interval for μ, when a **small** sample of size n is taken from a normal distribution (σ is unknown) is:

$$\overline{x} \pm t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Here, $t_{n-1,\alpha/2}$ is the critical value of the t-distribution with $n-1$ degrees of freedom.

Note: Provided that the population is reasonably close to being normally distributed (i.e. approximately normal, or near-normal), the above can still be used.

We only need to use the above for **small** sample sizes when σ is unknown. (If the sample size was large, then critical value from t_{n-1} and the critical value from Z would be nearly identical; if we knew σ, then our critical value would have been from Z.)

Example: The following data is collected on the mass (in grams) of adult white mice.

$$14.6, 13.2, 19.5, 10.1, 8.8, 15.5, 16.1$$

Assuming that the weights of mice are normally distributed, find a 95% confidence interval for μ, the mean weight of adult white mice.

Example: A random sample of nine observations on the number of pounds of nitrogen created per duck farm per day is taken:

$$4.9, 5.8, 5.9, 6.5, 5.5, 5.0, 5.6, 6.0, 5.7$$

Assuming that the number of pounds is approximately normally distributed, find a 99% confidence interval for μ.

Set 24

Suppose we make n observations from some binomial experiment and n is large. We would like to estimate the value of the true proportion of success, p.

If out of the n observations we have x successes, then $\widehat{p} = x/n$ is the **sample proportion**. We will use it as our point estimate for p.

The following has approximately standard normal distribution, regardless of the value of the parameter p.

$$\frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Note: $\sqrt{p(1-p)/n}$ is the **standard error** of $\widehat{p}$. Since the value of p is unknown (as we are currently estimating it), we can not use the standard error in our confidence interval.

Instead, we use the **estimated standard error** of $\widehat{p}$, which is $\sqrt{\widehat{p}(1-\widehat{p})/n}$.

The following is an approximate $100(1-\alpha)\%$ confidence interval for p:

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

Note: The validity of this depends on being able to approximate the binomial distribution with a normal distribution. Recall that to do so, we needed to have at least five successes and at least five failures.

Example: A sample of 1380 randomly selected books produced by a publishing company finds that 25 have bookbinding errors. Find a 95% confidence interval for p, the proportion of books with bookbinding errors.

If we are estimating sample size for population proportion, we'll find that for a given margin of error, d:

$$n = \frac{(z_{\alpha/2})^2 \widehat{p}(1-\widehat{p})}{d^2}$$

Note: To use this to calculate sample size, we need to know the value of $\widehat{p}$ before we carry out the study (which is of course impossible). In practice, we use a value from a prior study as an estimate for what $\widehat{p}$ will be.

If there has been no prior study, or if the prior estimate for p seems likely be different than the current value, there is another option:

We can show that $\widehat{p}(1-\widehat{p})$ has a maximum value of $1/4$, when $\widehat{p} = 1/2$. So the **maximum** sample size needed to guarantee a CI for p with a given margin of error is found by using:

$$n = \frac{(z_{\alpha/2})^2}{4d^2}$$

Example: In an earlier study, it was found that 1.4% of all microchips made by a particular manufacturer were defective. Using this as a pilot study, estimate the sample size needed to create a 99% confidence interval for p, the true proportion of defective microchips, with a margin of error of 0.005.

Example: We wish to carry out a telephone survey to estimate p, the proportion of island residents who want a bridge to the mainland. How many people must we call in order to estimate p with 98% confidence, to within 0.01?

Example: How many would we need to call in order to estimate p with 95% confidence, to within 0.025?