**An estimation problem**

Suppose we want to find the average height (in inches) of all male Calvin students.

**Population**: all male Calvin students

**Feature of interest**: height in inches

**Population Parameter**:  = mean/average height (population mean)

We could call up every male Calvin student, ask for his height, and average the results. This would give us the exact value of . That is great, but a lot of work. Also, in most cases, the population is too large to completely survey. So, we take an easier (and often the only practical) way. We take a random sample of 25 males from the population, compute the average height of that sample: = sample mean, and use  as an **estimate** of the actual average. The risk, of course, is that different samples will produce different values of , some of them close to  and some of them not close to . Since we do not know the value of , we have no way of telling whether our sample mean is “good” or not.

**Sample**: 25 individuals chosen (randomly) from the population

**Sample Statistic**: = sample mean

**Context**

Let’s suppose the heights of male Calvin students are normally distributed with mean  and standard deviation . If X is the height of a randomly selected male student, then X is a random variable with the distribution N(,). A random sample of size 25 is a set of 25 random variables X1,…,X25, of type N(,) that are independent. The sample mean defined on the samples is itself a RV, . The distribution of is the sampling distribution.

Sampling Distribution (of )

**Good news**  is an unbiased estimator for 

**Not so good news:** Some samples will give bad estimates.

How likely are we to get a good estimate? How likely is it that  will produce a value “close” to ? The smaller the sd(), more likely that we get a good estimate.

Suppose that  = 2.5 inches. Then,

So, for 95% of samples of size 25, the interval (-1, +1) will contain the true population mean . For a particular sample and sample mean , the interval ( is called a **95% confidence interval** for .

**In general,** if we know the population sd  and  is the mean of a random sample of size n, then

± 2 or, more precisely, ± 1.96

is a **95% confidence interval** for .

If my sample of 25 male Calvin students has a sample mean = 70.5, then the 95% CI for  is 70.5 ± 1 or (69.5,71.5).

What this doesn’t mean:

What this does mean:

Margin of Error

Other confidence levels:

90% ± 1.65

99% ± 2.58

**Note:** as the confidence level increases, the margin of error increases; i.e., the interval becomes less precise. Also, as the sample size increases, the margin of error of the CI decreases and the interval becomes more precise.

**What if we don’t know that the population we are sampling from is normal?**

Then, by Fact 2 for normal distributions, if n is “large”, the random variable  still is approximately normal, so  is approximately , and the confidence interval formulas above give us “good” approximate confidence intervals for , and the larger the sample size, the better the approximation.

Effect of large sample size:

1.

2.

**Example**: Let  be the mean weight in ounces of the contents of all 24 oz boxes of Kelloggs Frosted Miniwheats cereal. To estimate the value of , we measure the weight of the contents of a randomly selected sample of 100 boxes. Assume the standard deviation ,σ, of the weights of all boxes is 0.1 oz. If the sample mean of our sample is = 24.05 oz., compute the 95% CI for this sample.

**Back to the real world**

In practice, we rarely know what the standard deviation of the population is, so the method above doesn’t work. We follow the statistical maxim: if you need to know something and don’t have direct access to its value, estimate it from the data (sample). In this case, we can estimate the population standard deviation  using the sample standard deviation s. How good is s as an estimate of ? Like estimates in general, some samples give good estimates and some provide not-so-good estimates.

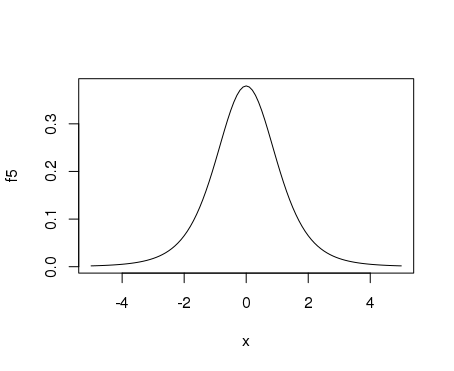
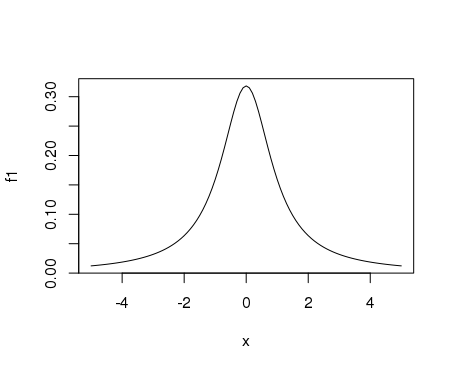
The **good news** is that S2, the sample variance random variable whose value depends on the sample, is an unbiased estimator of , the population variance.

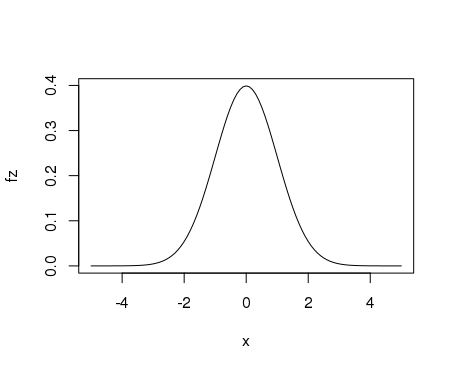
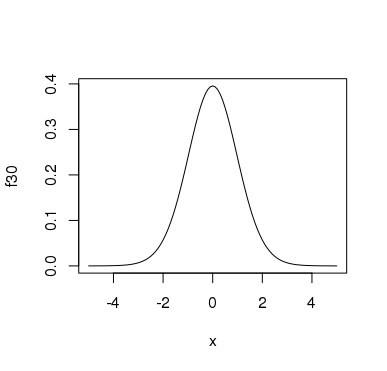
The **bad news** is that while  is standard normal,  is not normal. This fact was discovered by Gosset. The **good news** is that Gosset also discovered what the distribution of  is. This not normal random variable has a **t-distribution with (n-1) degrees of freedom, where n is the sample size.**

**The t distribution**

R has the t-distribution built in with the usual function forms: dt(x,n), pt(x,n),, qt(p,n), rt(k,n), where n = degrees of freedom

The t-distributions are symmetric, unimodal, with mean = 0. They have “fatter tails” that the standard normal distribution; i.e., there is more probability in the tails than with the standard normal distribution. Finally, as the value of n = df increases, the t-distribution gets closer to the standard normal distribution.





**t-based confidence intervals**

If the population is normal and and s are the mean and standard deviation of a random sample of size n, then the t-based confidence interval is ± tn-1\*, where tn-1\* depends on the confidence level.

**Example** If the sample size is n=10 and the confidence level is 95%, t9\* is the value for which 95% of the probability for a t-distribution with df =9 is in the range ± t9\*. Equivalently,97.5% of the probability is < t9\*. Using R:

> qt(.975,9)

[1] 2.262157

So t9\* for a 95% CI is 2.262.

In general:

for 95% CI, t\* = qt(.975,n-1)

for 99% CI, t\* = qt(.995,n-1)

for 90% CI, t\* = qt(.95,n-1)

**Example.** Suppose a sample of size n from a normal population has a mean  = 5 and standard deviation s = 2.

If n=15, find the 90% t-based confidence interval for the population mean .

> qt(.95,14)

[1] 1.76131

If n=50, find the 95% t-based confidence interval for mu.

> qt(.975,49)

[1] 2.009575

**Automating t-based confidence intervals using R**

**Example:** The dataframe **Dimes** contains the masses of a random sample of dimes. The column of masses is headed **mass.** Find the 90%, 95%, and 99% CI’s for the average mass of all dimes using this sample.

> confint(t.test(~mass,data=Dimes,conf.level=.90))

mean of x lower upper level

1 2.258233 2.251387 2.26508 0.9

> confint(t.test(~mass,data=Dimes,conf.level=.95))

mean of x lower upper level

2.258233 2.249992 2.266474 0.95

> confint(t.test(~mass,data=Dimes,conf.level=.99))

mean of x lower upper level

2.258233 2.247127 2.26934 0.99

**What if the underlying population isn’t normal? Robustness**

1. For normal distributions, t-based CI’s are always good, even for small samples.
2. For 15 ≤ n ≤ 40, t-based CI’s are acceptable if the distribution is unimodal and not strongly skewed.
3. For n > 40, t-based CI’s are generally acceptable if the distribution is unimodal.
4. Beware of outliers.

**Estimating population proportions**

**Problem**: We want to estimate the proportion of registered voters in Michigan that have a favorable view of Governor Whitmer The population parameter of interest is the population proportion p ( proportion of the population that have a favorable view of Snyder. )

**Solution**: We poll a random sample of 100 registered voters and measure the sample proportion.  = proportion of the sample with a favorable view of Whitmer. Is  a good estimate of p? We don’t know. Can we use  to construct a CI estimate for p? Yes. We need to know the sampling distribution for .

Let X be the random variable that has the value 1 for a registered voter with a favorable view and has the value 0 otherwise. On average, the value of X will be p for the population, so that we can think of p as a population mean. The sample will be a set of 0’s and 1’s: x1,…,x100 and =( x1+…+x100)/100 is the sample mean. The snag is that the population is definitely not a normal population, so there is no reason to expect that the sampling distribution of is a normal distribution. However, by Fact 2 of normal distribution (Central Limit Theorem) if the sample size n is large, ’s distribution is approximately normal (and the larger the sample size the better the approximation.)

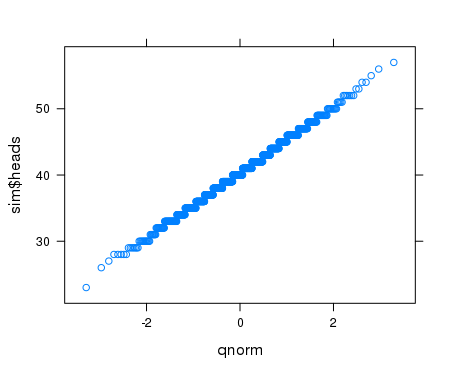
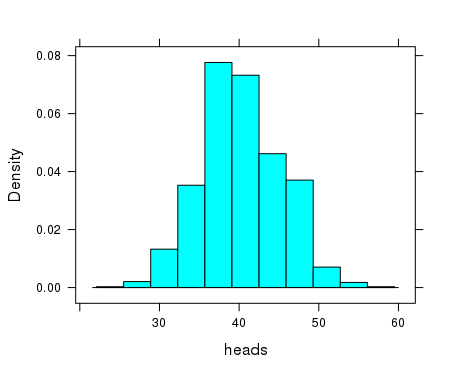
* sim<-do(1000)\*rflip(100,prob =.4)
* histogram(~heads, data=sim)

> qqmath(sim$heads)

> fitdistr(sim$heads,"normal")

mean sd

40.2950000 4.9799573



**Confidence intervals for population proportions**

If the sample size n is large, the sampling distribution for  is approximately normal and an approximate confidence interval for the population proportion p is ± z, where z depends on the confidence level.

90% z = 1.65

95% z = 1.96

99% z = 2.58

What is large enough?

**Example:** A poll taken at the end of August, 2017, of 600 registered Michigan voters produced 246 respondents that approved of Governor Snyder. Find the 95% CI for the proportion of all registered Michigan voters who approved of the governor August, 2017.

**Exercises 11**

1. A sample of size 100 produced  = 4.2 and s = 1.3. Find the 95% and the 99% confidence interval for the population mean .
2. The contents of 50 12 ounce cans of Coke Zero were measured. The average contents of those cans was 12.05 ounces with a standard deviation of 0.1 ounces.
3. Based on this data, compute the 95% CI for the average of all 12 ounce cans of Coke Zero.
4. Based on this confidence interval, are you confident that, on average, 12 ounce cans of Coke Zero contain at least 12 ounces of soda?
5. A sample of size 20 produced a sample mean of 3.2 and a sample sd of 1.1.
6. If it appears that the population the sample came from is strongly skewed to the right. Should you calculate a 95% CI using the t-distribution?
7. If it appears that the population is unimodal and quite symmetric, should you calculate a 99% CI using a t-distribution?
8. If a t-distribution is appropriate in (a) and/or (b), find the required CI.
9. Company B receives a large shipment of parts from Company A. Company B accepts the fact that any large shipment of parts will contain some defective parts. It is willing to accept a shipment that it is confident contains ar most 10% defective parts. It examines 200 randomly selected parts from the shipment and finds that 15 of them are defective. Calculate the 95% CI for the proportion of all parts in the shipment that are defective. Based on this CI, should Company B be confident that the defective rate for the entire shipment is at most 10%?
10. The data-frame morley contains measurement of the speed of light in the column Speed. We can consider this data to be a random sample of size 100 of all possible measures of the speed of light. Use this data to find a 95% CI for the average of all possible measurements.
11. A recent Gallup poll of 3500 people produced a approval rating for President Obama of 50%. Assuming the confidence level is 95%, what is the margin of error?