Chapter 6

Point Estimation and Sampling Distributions

Point Estimation

As discussed in the previous section we use statistics to estimate **population parameters**. When we estimate a **target parameter** with single value we call it a **point estimate**. $\overline{x}$ is a point estimate for μ, and $\hat{p}$ is a point estimate for p.

Bias of a Point Estimator: We say $\hat{\theta}$ is an **unbiased estimator** of population parameter θ if $E(\hat{\theta}) = \theta$.

We denote the bias of a point estimator $\hat{\theta}$ as $B(\theta)$.

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

We are also interested in the variance of estimators, and how they are distributed.

What is a Sampling Distribution?

A sampling distribution refers to the distribution that is formed for a statistic over all possible samples. Suppose you measure a statistic from a random sample. If you were to conduct the exact same random sample ad infinitum the distribution of all statistics from all samples form the sampling distribution. Sampling distributions are absolutely instrumental for statistical inference.

The Sampling Distribution for $\hat{p}$

Let us first consider how the sample proportion is calculated. A random sample of n elements is gathered from a population of N. The number of objects that fall in a particular category are counted, we will denote this total X.

$$\hat{p} = \frac{X}{n}$$

First let's consider the actual distribution of X. We are selecting n objects from a population of N, and counting the number of sample points in n have some particular characteristic. r denotes the the total number of objects that are classified as 'successes' in the population.

This means that the population proportion $p = \frac{r}{N}$. Recall from Chapter Two that this scenario describes a hypergeometric random variable. We have $X \sim \text{Hypergeometric}(r, n, N)$. This allows us to find the expected value of $\hat{p}$

$$\begin{aligned} E(\hat{p}) &= E\left(\frac{X}{n}\right) \\ &= \frac{1}{n}E(X) \\ &= \frac{1}{n}n\left(\frac{r}{N}\right) \\ &= \frac{r}{N} \\ &= p \end{aligned}$$

Notice that the expected value for $\hat{p}$ is p, the true population proportion. $\hat{p}$ is an unbiased estimator of p. If we conduct the same sample over and over again, our sampling distribution for $\hat{p}$ will be centered at the true population proportion.

We may also find the variance and standard deviation of $\hat{p}$

$$\begin{aligned} Var\left(\hat{p}\right) &= Var\left(\frac{X}{n}\right) \\ &= \frac{1}{n^2}n\left(\frac{r}{N}\right)\left(1-\frac{r}{N}\right)\left(\frac{N-n}{N-1}\right) \\ &= \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right) \\ SD(X) &= \sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)} \end{aligned}$$

This is most appropriate for small finite populations. The hypergeometric distribution calculations quickly get out of hand for large populations. For large populations we will use an approximate distribution. We will assume that $X \sim \text{Binomial}(n, p)$, but this assumption is only appropriate under certain conditions.

- **Independence Condition:** We require that $N >> n$, so that the probability of selecting a 'success' between sample points is equal. Our 'cutoff' for this condition will be $n < 10\%$ of N.

Now, assuming $X \sim \text{Binomial}(n, p)$ we can find the mean and variance:

$$\begin{aligned} E(\hat{p}) &= E\left(\frac{X}{n}\right) \\ &= \frac{1}{n}E(X) \\ &= \frac{1}{p}(np) \\ &= p \end{aligned}$$

$$\begin{aligned} Var(\hat{p}) &= Var\left(\frac{X}{n}\right) \\ &= \frac{1}{n^2}Var(X) \\ &= \frac{1}{n^2}(np(1-p)) \\ &= \frac{p(1-p)}{n} \\ SD(X) &= \sqrt{\frac{p(1-p)}{n}} \end{aligned}$$

Notice that $\hat{p}$ is still unbiased in estimating p. But another problem arises; with large numbers the calculations also get out of hand for a binomial random variable so we would like to approximate using the normal distribution. As discussed earlier this is only appropriate for a sufficiently large n and p.

- **Normality Condition:** The sampling distribution for $\hat{p}$ is approximately normal if $np > 10$, and $n(1-p) > 10$. This can also be seen as 10 'successes' and 10 'failures' in the sample.

Lastly, we must ensure every sample points are truly behaving as random variables.

- **Random Sampling Condition:** Samples points must be drawn using random sampling.

These conditions are sometimes ambiguous to check concretely. We must assume them to be true in order to use this sampling distribution. If we are able to make these assumptions we have

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim Z$$

For an introductory statistics course we will only interest ourselves in the case where $\hat{p}$ is normal. So we will require **random sampling**, **independence** and **normality**.

Example 1: It is known that across North America 65% of University students take longer than four years to complete their undergraduate degree. You conduct a survey of 100 University of Calgary graduates and ask them if it took longer than four years to complete their degree.

(a) Let X be the number of students in your sample that took longer than four years to complete their degree. What is the distribution for X?

Here we have $X \sim$ Hypergeometric. However, as n is clearly less than 10% of N we will say $X \sim \text{Binomial}(n, p)$. Where n is the sample size and p is the true proportion of college graduates who took longer than four years to graduate. Here $n = 100$, and $p = 0.65$.

(b) What is the sampling distribution for $\hat{p}$, the proportion of students in the sample who took longer than 4 years to complete their degree?

First we will notice that the independence condition if fulfilled as $n < 0.1N$. This sample also satisfies the normality condition as $n(p) = 65$ and $n(1-p) = 35$. We have $\hat{p} \sim \text{Normal}\left(p, \frac{p(1-p)}{n}\right)$.

(c) What is the probability that the sample proportion is larger than 70% in your sample?

Here we would solve directly using the sampling distribution for $\hat{p}$, but we will standardize. We know that

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim Z$$

Where $Z \sim \text{Normal}\,(0,1)$. We have

$$\begin{aligned} z &= \frac{0.70-0.65}{\sqrt{\frac{0.65(0.35)}{100}}} \\ &= 1.048285 \end{aligned}$$

And then we may determine probability using a computer or tables.

$$\begin{aligned} P(\hat{p} > 0.70) &= P(Z > 1.048285) \\ &= 1 - P(Z \leq 1.048285) \\ &= 0.1472537 \end{aligned}$$

So there is an approximately 15% chance that the sample will be comprised of 70% or more students who took longer than 4 years to finish their under graduate degree.

The Sampling Distribution for $\overline{x}$

Now suppose X_1, X_2, ... , X_n are independent random variables from the same 'parent distribution'. The sample mean, $\overline{x}$ is defined by

$$\overline{x} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

First we will consider the case where the 'parent distribution' is **normally distributed**. In other words $X_i \sim \text{Normal}(\mu, \sigma^2)$ for $i = 1, 2, 3, \ldots, n$. Let's first look at the expected value, variance and standard deviation for $\overline{x}$.

$$\begin{aligned} E(\overline{x}) &= E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\ &= \frac{1}{n}E(X_1 + X_2 + \cdots + X_n) \\ &= \frac{1}{n}\left(E(X_1) + E(X_2) + \cdots + E(X_n)\right) \\ &= \frac{1}{n}\left(\mu + \mu + \cdots + \mu\right) \\ &= \frac{n\mu}{n} \\ &= \mu \end{aligned}$$

So the sample mean $\overline{x}$ is an unbiased estimator for the population mean μ.

$$\begin{aligned} Var(\overline{x}) &= Var\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\ &= \frac{1}{n^2}Var(X_1 + X_2 + \cdots + X_n) \\ &= \frac{1}{n^2}\left(Var(X_1) + Var(X_2) + \cdots + Var(X_n)\right) \\ &= \frac{1}{n^2}\left(\sigma^2 + \sigma^2 + \cdots + \sigma^2\right) \\ &= \frac{n\sigma^2}{n^2} \\ &= \frac{\sigma^2}{n} \\ SD(\overline{x}) &= \frac{\sigma}{\sqrt{n}} \end{aligned}$$

From earlier chapters we know that the distribution for a sum of normal random variables is also a normal random variable. So we have $\overline{x} \sim \text{Normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. As each sample point must be independent we also require that the sample is less than or equal than 10 percent of the population size. This is the **independence condition**; $n < 10\%$ of N. We often prefer the standardized version of the normal distribution.

$$\frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim Z$$

Notice that this distribution requires knowing σ. We will need a new approach if σ is unknown. This will be covered later in the chapter.

Example 1: The weights of pale-throated sloths are known to follow a normal distribution with a mean weight of 4.5 kg, and standard deviation of 1.1 kg. Suppose that you randomly sample 20 sloths. What is the probability that your samples has an average weight of 4.8 kg or less?

(a) Describe the sampling distribution for $\overline{x}$.

Here we have a normal parent distribution and may also assume than $n = 20$ is less than 10% of the total population size of all sloths ($n < 0.1N$). So we have $\overline{x} \sim \text{Normal}\left(\mu = 4.5, \frac{\sigma^2}{n} = \frac{(1.1)^2}{20}\right)$.

(b) What is the probability that your sample has an average of between 2.3 kg and 4.3 kg?

$$z_{low} = \frac{2.3 - 4.5}{\frac{1.1}{\sqrt{20}}} = -8.944272$$

almost 9 standard deviations below the mean!

$$z_{high} = \frac{4.3 - 4.5}{\frac{1.1}{\sqrt{20}}} = -0.8131156$$

$$\begin{aligned} P(2.3 \leq \overline{x} \leq 4.3) &= P(-8.9 \leq Z \leq -0.8) \\ &= P(Z < -0.8) - P(Z < -8.9) \\ &= 0.2118554 \end{aligned}$$

There is an approximately 21.2% chance of observing an average of between 2.3 kg and 4.3 kg.

(c) What is the probability a randomly selected sloth has a weight greater than 4.6 kg?

Careful! We aren't talking about the sampling distribution here. We are simply drawing a random point from the parent distribution. Let X be the sloth weights. $X \sim \text{Normal}(4.5, 1.1)$. We have

$$\begin{aligned} P(X > 4.6) &= P\left(Z > \frac{4.6 - 4.5}{1.1}\right) \\ &= P(Z > 0.09090909) \\ &= 1 - P(Z \leq 0.09090909) \\ &= 0.4637824 \end{aligned}$$

There is a roughly 46% chance of observing a pale-throated sloth with a weight greater than 4.6 kg.

The Central Limit Theorem

We are now comfortable finding the sampling distribution for $\overline{x}$ if the parent distribution is normal. But what if you are sampling from a population with a non-normal distribution? Now we will introduce the **central limit theorem**.

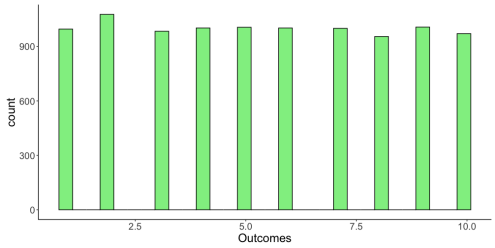
The Central Limit Theorem: Let X_1, X_2, ... , X_n be independent identically distributed random variables (i.i.d) with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. The sample mean is defined by

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

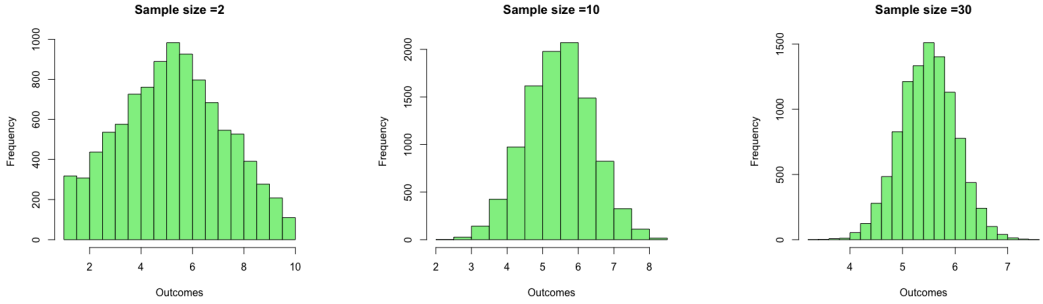
As n tends to infinity, the sampling distribution for $\overline{X}$ converges to a normal distribution with mean $\mu_{\overline{x}} = \mu$ and variance $\sigma^2_{\overline{x}} = \frac{\sigma^2}{n}$. In other words as $n \to \infty$, $\overline{X} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$, regardless of the parent distribution. We will assume samples of size $n \geq 30$ are large enough for the central limit theorem to apply. We may also standardize our result; as n gets sufficiently large ($n \geq 30$).

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim Z$$

Example 1: Suppose you are rolling a 10 sided die. Every outcome has an equal chance of occurring. Shown below is the probability distribution for 10000 rolls of the die. The probabilities are roughly uniform for all outcomes. This is the 'parent' distribution. Let's let X denote the number that appears on each roll of the dice.



Next we can look at the sampling distribution for samples of different sizes. Here X_1, X_2,..., X_n represent the outcomes for the sample of n die. Show below are the sampling distributions of $\overline{X}$ for 10000 samples of size 2, 10, and 30.



We can see that as the sample size increases, the sampling distribution for $\overline{X}$ converges to a normal distribution.

Example 2: A certain carnival game is designed with the following profit distribution

X (Profit)	-\$1	\$1	\$5	\$20
$P(X = x)$	0.95	0.03	0.02	0.01

(a) What is a player's expected earnings?

$$\begin{aligned} E(X) &= -1(0.95) + 1(0.03) + 5(0.02) + 20(0.01) \\ &= -0.62 \end{aligned}$$

(b) Determine $Var(X)$.

$$\begin{aligned} E(X^2) &= (-1)^2(0.95) + (1)^2(0.03) + (5)^2(0.02) + (20^2)(0.01) \\ &= 5.48 \end{aligned}$$

$$\begin{aligned} Var(X) &= E(X^2) - E(X)^2 \\ &= 5.48 - (-0.62)^2 \\ &= 5.4556 \end{aligned}$$

(c) Suppose each night you visit the carnival you play the game 30 times. Describe the distribution for your average nightly earnings.

Here we have a large enough sample for the central limit theorem to apply. We have $\mu = -0.62$ and $\frac{\sigma^2}{n} = \frac{5.4556}{30} = 0.1818533$. This suggests that $\overline{x} \sim \text{Normal}(\mu_{\overline{x}} = -0.62, \sigma^2_{\overline{x}} = 0.18)$. We also have $\sigma_{\overline{x}} = \sqrt{0.1818533} = 0.4264426$

(d) What is the probability that your average nightly profit is positive?

$$\begin{aligned} P(\overline{x} > 0) &= P\left(\frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} > \frac{0 - (-0.62)}{\frac{0.4264426}{\sqrt{30}}}\right) \\ &= P(Z > 7.963275) \\ &= 1 - P(Z \leq 7.963275) \\ &\approx 0 \end{aligned}$$

(e) What's the probability you win more than \$1 when playing the game?

Careful! We aren't talking about the sampling distribution here, but rather the parent distribution. By simple inspection of the distribution we have $P(X > 1) = P(X = 5) + P(X = 20) = 0.03$

Assumptions For Using A Normal Model
To summarize, the normal model may only be used for the sampling distribution of $\overline{x}$ under certain conditions. In practice we must assume these conditions hold true

- **Normality:** We require the sample is greater than 30, so the central limit theorem applies **or** we require the parent population to be normal.
- **Independence:** We require sample points to be independent of each other. We may make this assumption whenever $n < 10\%$ of N.
- **Random Sampling:** We require that samples are drawn using a random sampling technique, so we may treat each sample point as a random variable.

When all assumptions can be made we have $\overline{x} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$. Notice that the assumptions we have to make are very similar to that of the sample proportion.

A Sampling Distribution involving s^2

Next we will discuss another sampling distribution involving the sample variance s^2, and population variance σ^2. Assume X_1, X_2, ... , X_n are a random sample of size n from a normal distribution with mean μ and variance σ^2.

$$\frac{s^2(n-1)}{\sigma^2} \sim \chi^2_{n-1}$$

Example 1: The weights of pale-throated sloths are known to follow a normal distribution with a mean weight of 4.5 kg, and standard deviation of 1.1 kg. Suppose that you randomly sample 20 sloths. What is the probability that your sample has a standard deviation of 0.9 or greater?

Notice that σ and s^2 are both known in this problem.

$$\begin{aligned} P(s^2 > 0.9^2) &= P\left(\frac{s^2(n-1)}{\sigma^2} > \frac{(0.9)^2(20-1)}{(1.1)^2}\right) \\ &= P\left(\chi^2_{19} > 12.71901\right) \\ &= 1 - P\left(\chi^2_{19} \leq 12.71901\right) \\ &= 0.8526398 \end{aligned}$$

There is an approximately 85% chance of observing a standard deviation of 0.9 or greater in the sample.

6.6.1 Standard Deviation and Standard Error

In some of the distributions we have examined so far we run into a slight issue; the standard deviation of the sampling distribution is a function of the population parameters. For example

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

This is a problem because population parameters are often unknown. In this case we want to estimate the parameter with a statistic. We may estimate the population standard deviation σ with the sample standard deviation s. We call this estimate the **standard error**. This will be an instrumental idea when constructing confidence intervals in upcoming sections. In general, standard error is a term used to describe the estimated standard deviation of a statistic.

But we run into another issue; when we estimate a parameter with a statistic the distribution can change. For example, after making the appropriate assumptions we know

$$\frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim Z$$

But after estimating σ with s, what is the new distribution?

$$\frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} \sim ?$$

Using the t Distribution

Consider a standard normal random variable Z, and χ^2 random variable with k degrees of freedom. The t distribution is defined as

$$t = \frac{Z}{\sqrt{\frac{\chi^2_k}{k}}}$$

We also know that in certain circumstances (recall conditions) that

$$\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right) \sim Z \quad \text{and} \quad \frac{s^2(n-1)}{\sigma^2} \sim \chi^2_{n-1}$$

Using these statistics we can construct a t distribution

$$\frac{\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right)}{\sqrt{\frac{\left(\frac{s^2(n-1)}{\sigma^2}\right)}{(n-1)}}} = \frac{(\overline{X} - \mu)}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

This is the case where we have estimated σ using s.

Now let's discuss the assumptions that must be made when using the t distribution.

- **Simple Random Sampling:** We require our sample to be drawn using a random sampling technique. This ensures $X_1, X_2, \ldots, X_n$ are truly random variables.
- **Independence:** We require that $X_1, X_2, \ldots, X_n$ are independent random variables. For sampling without replacement we will require $n < 10\%N$ so that each sample point is independent.
- **Normality:** We require $X_1, X_2, \ldots, X_n$ come from a normal population with mean μ, and variance σ^2. This ensures that

$$\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right) \sim Z \quad \text{and} \quad \frac{s^2(n-1)}{\sigma^2} \sim \chi^2_{n-1}$$

However, due to the central limit theorem we know for samples of $n \geq 30$ we will have

$$\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right) \sim Z$$

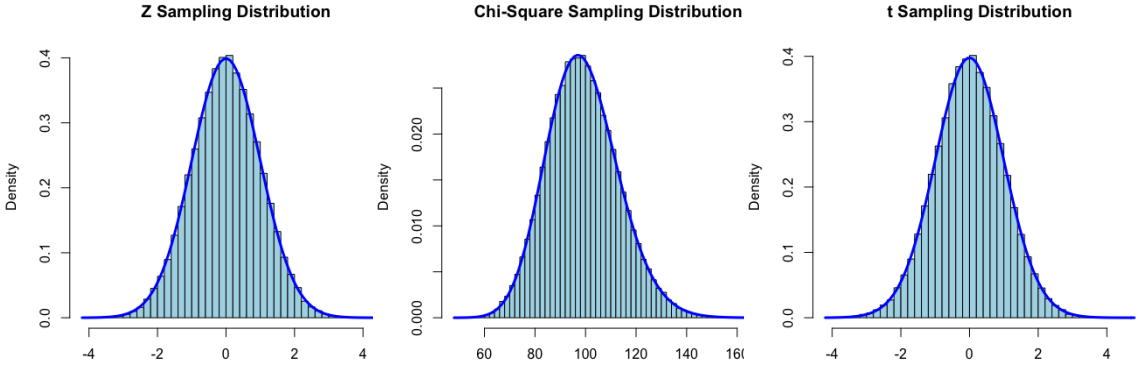
It turns out the t distribution is quite robust. If the parent population is roughly normal or at least symmetrical, unimodal, and without outliers and we have a sufficiently large sample it's alright to assume that

$$\frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

We must be *extremely* cautious when dealing with small samples of data that appears non-normal, heavily skewed, or heavy tailed. A t distribution may be *entirely* inappropriate.

Let's illustrate this idea by observing several cases.

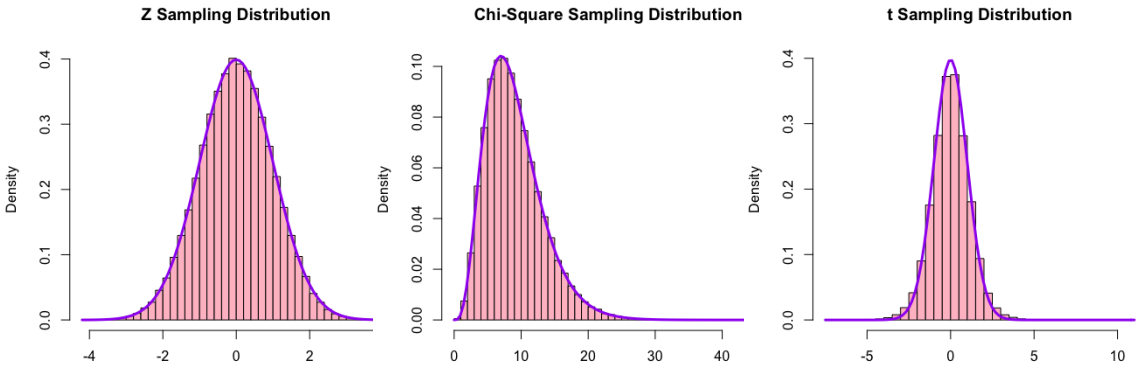
Case I→ Let's assume that $X_1, X_2, \ldots, X_{100}$ are independent, identically distributed variables with $X_i \sim$ Normal(10,2). Let's look at the sampling distribution for some of the statistics we discussed over 100000 generated samples.



$$\frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim Z \qquad \frac{s^2(n-1)}{\sigma^2} \sim \chi^2_{99} \qquad \frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{99}$$

The histograms are showing the experimental sampling distribution for 100,000 samples. The superimposed curve is the theoretical sampling distribution. This is our ideal scenario, we have a large sample size, and a normal parent population. You can see that under the correct conditions it is quite a nice fit.

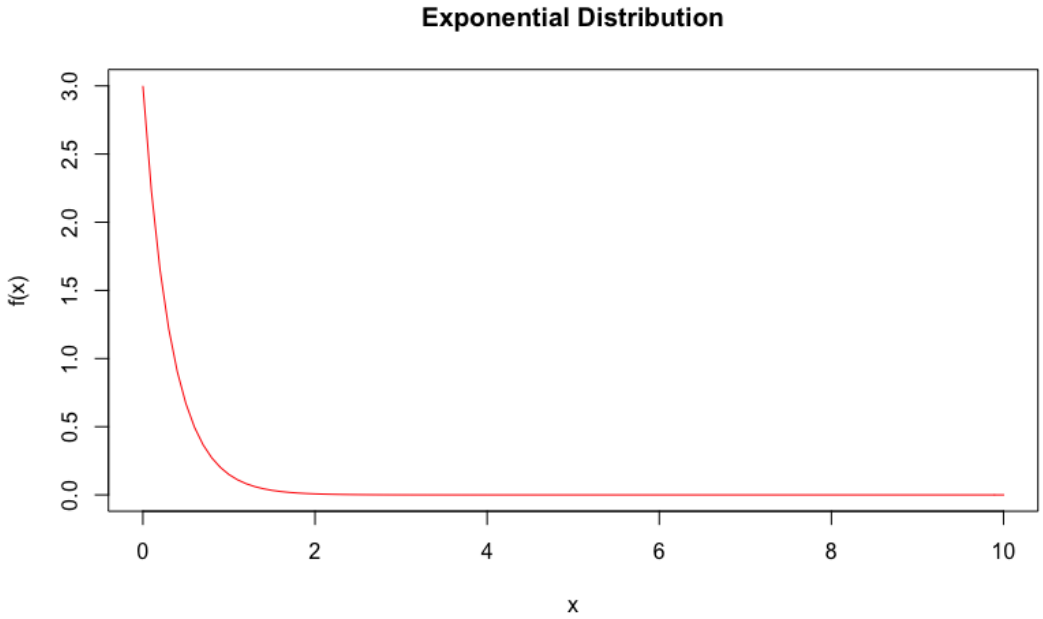
Case II→ Now consider a sample of X_1, X_2, …, X_{10} where $X \sim \text{Normal}(10, 2)$. This is the same as case one, but with a small sample size. Let's inspect the distribution.



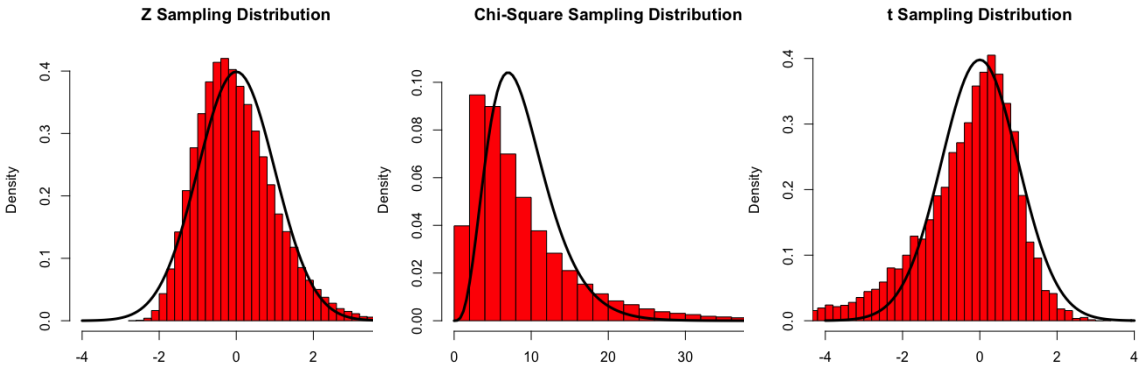
$$\frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim Z \qquad \frac{s^2(n-1)}{\sigma^2} \sim \chi^2_9 \qquad \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} \sim t_9$$

Notice that since we have normality in the parent distribution, the t distribution is an appropriate fit regardless of the small sample size.

Case III→ Let's assume that X_1, X_2, …, X_{10} are identically distributed variables with $X_i \sim \text{Exp}(3)$. **Yikes!** We haven't explored the exponential distribution at all this far, however the pdf is shown below:



The important takeaway here is that the distribution is nowhere near normal! It's skewed to the right. Another cause for concern is our sample is only comprised of 10 points! Let's look at the sampling distribution for some of the statistics we discussed over 100000 generated samples.



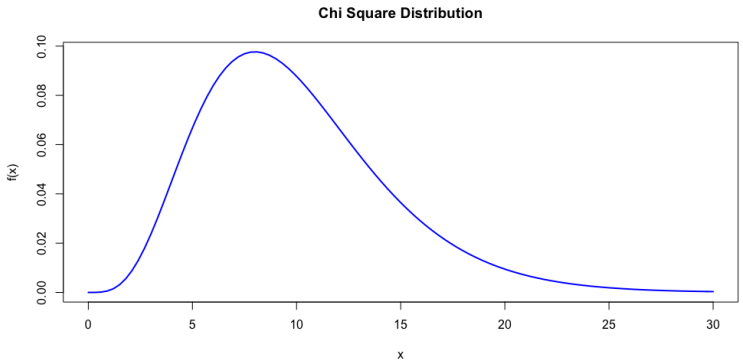
$$\frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim ? \qquad \frac{s^2(n-1)}{\sigma^2} \sim ? \qquad \frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}} \sim ?$$

You can see with the skewed parent population and low sample size the theoretical sampling distributions are a terrible fit. This is why we must be extremely mindful of the assumptions we are making when using the t distribution.

The important takeaway here is that the t distribution is really only appropriate under certain circumstances. Let's reiterate

- **Simple Random Sampling:** We require our sample to be drawn using a random sampling technique. This ensures $X_1, X_2, \ldots, X_n$ are truly random variables.
- **Independence:** We require that $X_1, X_2, \ldots, X_n$ are independent random variables. For sampling without replacement we will require $n < 10\%N$ so that each sample point is independent.
- **Normality:** We require that $n \geq 30$ **or** that the parent population is approximately normal. Even with relatively large sample sizes we should still be weary of skewness and heavy tails in data.

Example 1: $X_1, X_2, \ldots, X_{500}$ are independent random variables that form a random sample of $n = 500$ with $X_i \sim \chi^2_{10}$ for $i = 1, 2, 3, \ldots, 500$. $f(x)$, the probability density function for the parent distribution is shown below



1. Describe the probability distribution for $\overline{x}$.

   Recall that for a Chi Square random variable X with k degrees of freedom we have $E(X) = k$, and $Var(X) = 2k$. In this example $k = 10$. Since we have a large sample size $n = 500$ we will have $\overline{x} \sim \text{Normal}\left(10, \sqrt{\frac{20}{500}}\right)$.

2. Find the probability that the sample mean will be greater than 9.8 in future samples

   When we have previously defined the sampling distribution this is easy.

$$\begin{aligned} P(\overline{X} > 9.8) &= P\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < \frac{9.8-10}{\sqrt{\frac{20}{500}}}\right) \\ &= P(Z > -1) \\ &= 1 - P(Z \leq -1) \\ &= 0.8413447 \end{aligned}$$

3. Find the probability that the sample mean will be between 0 and 10.1 in future samples

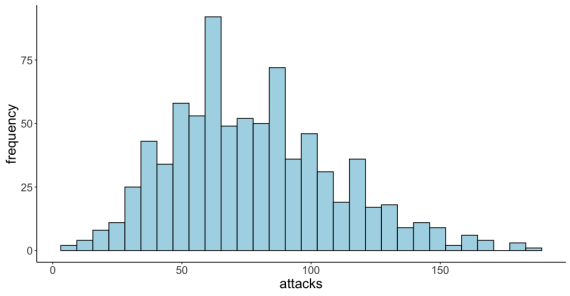
$$\begin{aligned} P\left(0 \leq \overline{X} \leq 10\right) &= P\left(\frac{0-10}{\sqrt{\frac{20}{500}}} \leq \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{10.1-10}{\sqrt{\frac{20}{500}}}\right) \\ &= P(-50 \leq Z \leq 0.5) \\ &= P(Z \leq 0.5) - P(Z \leq -50) \\ &= 0.6914625 \end{aligned}$$

4. Determine $P(X_1 > 1)$

   Be careful here. We are just talking about the parent distribution. This has nothing to do with the sample mean.

$$\begin{aligned} P(X_1 > 10) &= P(\chi^2_{10} > 10) \\ &= 1 - P(\chi^2_{10} \leq 10) \\ &= 0.4404933 \end{aligned}$$

Example 2: A random sample of $n = 801$ Pokémon has an average attack score of $\overline{x} = 78$, with standard deviation $s = 32$. Suppose that the true average attack of Pokémon is know known to be $\mu = 70$. The distribution for the sample is shown below.



(a) What is the probability that you will find an average score that is less than 75 in future samples?

   Here we have a large sample $n = 801$, and relatively normal data, so we may use a t distribution.

$$\begin{aligned} P(\overline{x} < 78) &= P\left(\frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}} < \frac{70-78}{\frac{32}{\sqrt{801}}}\right) \\ &= P(t_{800} < -2.653307) \\ &= 0.004064423 \end{aligned}$$

6.7.1 Quick Review

let's quickly summarize the results that we have formulated so far

Distribution	Assumptions
$\hat{p} \sim \text{Normal}\left(p, \frac{p(1-p)}{n}\right)$	**- Random Sampling** **- Independence** **- Normality**
$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim Z$	**- Random Sampling** **- Independence** **- Normality**
$\overline{x} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$	**- Random Sampling** **- Independence** **- Normality**
$\frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim Z$	**- Random Sampling** **- Independence** **- Normality**
$\frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$	**- Random Sampling** **- Independence** **- Normality**

A Difference of Proportions

We are also interested in the sampling distribution of the difference between two sample proportions for two different samples. Suppose we have two samples of size n_1 and n_2 with $\hat{p}_1$, and $\hat{p}_2$. In order to arrive at our sampling distribution we have a few requirements:

- **Random Sampling:** We require that both samples are drawn using random sampling. We also require that samples are independent.
- **Independence:** We require independence between each observation in our sample. We can make this assumption whenever $n_1 < 10\%$ of N_1 and $n_2 < 10\%$ of N_2.
- **Normality:** Just like the one sample case, we would like to approximate our sampling distribution using a normal distribution. This will require large enough samples. This assumption can be made whenever $n_1 p_1 > 10$, $n_1(1-p_1) > 10$, $n_2 p_2 > 10$, and $n_2(1-p_2) > 10$. This can also be seen as 10 'successes' and 10 'failures' in each sample.

Let's look at the mean and variance for a difference in proportions. flow

$$\begin{aligned} E(\hat{p}_1 - \hat{p}_2) &= E(\hat{p}_1) - E(\hat{p}_2) \\ &= p_1 - p_2 \end{aligned}$$

$$\begin{aligned} Var(\hat{p}_1 - \hat{p}_2) &= Var(\hat{p}_1) + Var(\hat{p}_2) \\ &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \end{aligned}$$

$$SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

If we are able to make all the appropriate assumptions we have

$$\hat{p}_1 - \hat{p}_2 \sim \text{Normal}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

We may also standardize this distribution

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim Z$$

Example 1: In one town 51% of the voters are conservative and in a second town 44% of the voters are conservative. Suppose 100 voters are surveyed from each town.

(a) Is a normal model appropriate for the difference in proportions of conservative voters from the two samples?

We need to check our requirements for the normal model

- **Simple Random Sample:** here we will assume the samples are drawn randomly
- **Independence:** We will assume both towns have more than 1000 people, so $n_1 < 10\%N_1$ and $n_2 < 10\%N_2$. This means we will assume sample points are independent.
- **Normality:** We have 51 expected successes and 49 expected failures in the first town, and 44 expected successes and 56 expected failures in the second town. This means we can assume there will be normality in the sampling distribution for the difference of proportions.

(b) What is the probability that the first sample will yield a lower sample proportion of conservative voters than the second town?

$$\begin{aligned} P\left(\hat{p}_1 - \hat{p}_2 < 0\right) &= P\left(\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} < \frac{0 - (.51 - .44)}{\sqrt{\frac{(.51)(0.49)}{100} + \frac{(0.44)(0.56)}{100}}}\right) \\ &= P(Z < -0.9936328) \\ &= 0.1602008 \end{aligned}$$

6.8.1 Populations with a Common Proportion

Sometimes the population proportion may be same for two different populations; $p_1 = p_2 = p_c$. As long as our requirements are fulfilled we will have

$$\hat{p}_1 - \hat{p}_2 \sim \text{Normal}\left(0, \frac{p_c(1-p_c)}{n_1} + \frac{p_c(1-p_c)}{n_2}\right)$$

$$\frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{p_c(1-p_c)}{n_1} + \frac{p_c(1-p_c)}{n_2}}} \sim Z$$

A Difference of Means

As with proportions, we are also interested in the sampling distribution that formed from the difference between two sample means, $\overline{x}_1 - \overline{x}_2$. In order to use this sampling distribution we have several requirements.

- **Random Sampling:** We require that both samples are drawn using random sampling.
- **Independence:** We require independence between each observation in our sample. We can make this assumption can be made whenever $n_1 < 10\%$ of N_1 and $n_2 < 10\%$ of N_2.
- **Normality:** Just like the one sample case, we would like to approximate our sampling distribution using a normal distribution. Here we will require $n_1 \geq 30$, and $n_2 \geq 30$, or both samples come from normal parent populations (data that appears approximately normal).

Just like with proportions we require that the samples are independent of each other. Let's take a look at the mean and variance.

$$\begin{aligned} E(\overline{x}_1 - \overline{x}_2) &= E(\overline{x}_1) - E(\overline{x}_2) \\ &= \mu_1 - \mu_2 \end{aligned}$$

$$\begin{aligned} Var(\overline{x}_1 - \overline{x}_2) &= Var(\overline{x}_1) + Var(\overline{x}_2) \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \end{aligned}$$

$$SD(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

If we are able to make all appropriate assumptions then we have

$$\overline{x}_1 - \overline{x}_2 \sim \text{Normal}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

We may also standardize this distribution

$$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim Z$$

Example 1: The starship Enterprise is exploring a new planet. Spock is inspecting two different alien species. Suppose the first species has a mean weight of 5 kg with a standard deviation of 1 kg, and the second species has a mean weight of 6 kg with a standard deviation of 2 kg. Spock randomly samples 63 of species one, and 80 of species 2.

(a) Is a normal model appropriate for a difference in sample means?

Let's look at the requirements for the normal model:

- **Random Sampling:** It is clearly stated that Spock takes a simple random sample
- **Independence:** The population sizes for the two types of aliens are not clearly stated but we will assume that $n_1 < 10\%N_1$ and $n_2 < 10\%N_2$.
- **Normality:** Here we have a sample of $n_1 = 62$ and $n_2 = 80$.

As the sample fulfills all requirements the normal model is appropriate.

(b) What is the probability that species one is greater than 1 kg heavier than species two?

$$\begin{aligned} P\left(\overline{x}_1 - \overline{x}_2 > 1\right) &= P\left(\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > \frac{1 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) \\ &= P\left(Z > \frac{1 - (5-6)}{\sqrt{\frac{1^2}{63} + \frac{2^2}{80}}}\right) \\ &= P(Z > 7.792489) \approx 0 \end{aligned}$$

6.9.1 Revisiting the t Distribution: Two Samples

Much like the one sample case σ_1 and σ_2 are often unknown to a researcher. In this case we will use a familiar strategy. We will estimate σ_1 with s_1 and σ_1 with s_2. Now we will examine the distribution of

$$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Next will divide this scenario into two cases:

- **Case I:** $\sigma_1^2 = \sigma_2^2$, the two populations share a common standard deviation. We will refer to this as the **'pooled variance'** case.
- **Case II:** $\sigma_1^2 \neq \sigma_2^2$, the two populations do not share a common standard deviation. We will call this the **'non-pooled variance'** case.

6.9.2 The Pooled Variance Case

When we assume two populations have the same variance, $\sigma_1 = \sigma_2$, we may estimate the common variance with the pooled sample variance.

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

After we estimate both s_1^2 and s_2^2 with s_p^2 we arrive at our result

$$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \sim t_{n_1+n_2-2}$$

Notice that the degrees of freedom aligns with the formula for the pooled variance. We can construct this distribution using the same process as the one sample case.

Let's look at the assumptions we must make for these results to hold true.

- **Equal variances:** We require the two populations to have a common variance that we estimate using pooling. We call two populations with the same variance **homoscedastic**.
- **Random Sampling:** We require both samples to be drawn randomly.
- **Independence:** We require that each sample is less than 10% of its respective population
- **Normality:** We require that the two parent populations are approximately normal or the sample sizes are sufficiently large ($n_1 \geq 30$, $n_1 \geq 30$).

6.9.3 The Non-pooled Variance Case

In this scenario we cannot rely on the assumption that $\sigma_1^2 = \sigma_2^2$. With $\sigma_1^2 \neq \sigma_2^2$, we must estimate σ_1^2 with s_1^2, and σ_2^2 with s_2^2. Looking at the sampling distribution we have

$$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{df}$$

Finding the degrees of freedom for this sampling distribution is less intuitive. The derivation is left for a more advanced course.

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

To ensure this sampling distribution is appropriate we must make several assumptions

- **Non Equal Variances:** We are assuming the two populations do not have a common variance. We call two populations with different variances **heteroscedastic**.
- **Random Sampling:** We require both samples to be drawn randomly.
- **Independence:** We require that each sample is less than 10% of its respective population
- **Normality:** We require that the two parent populations are approximately normal or the sample sizes are sufficiently large ($n_1 \geq 30$, $n_2 \geq 30$).

Notice that in both of these cases we introduce some ambiguity. How 'normal' must a sample be in order to assume normality in the parent population? How do we tell if two variances are statistically different? We will examine these questions more deeply when we explore hypothesis testing.

Let's summarize the 'difference' distributions we have looked at.

Distribution	Assumptions
$\hat{p}_1 - \hat{p}_2 \sim \text{Normal}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$	**- Random Sampling** **- Independence** **- Normality**
$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim Z$	**- Random Sampling** **- Independence** **- Normality**
$\overline{x}_1 - \overline{x}_2 \sim \text{Normal}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$	**- Random Sampling** **- Independence** **- Normality**
$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim Z$	**- Random Sampling** **- Independence** **- Normality**
$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \sim t_{n_1+n_2-2}$	**- Equal Variances** **- Random Sampling** **- Independence** **- Normality**
$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}} \sim t_{df}$	**- Nonequal Variances** **- Random Sampling** **- Independence** **- Normality**