Samples with Normally Distributed Data

We have talked about using the binomial as a sampling distribution for data that are categorical with two categories. And we can use this sampling distribution to estimate the fraction of the population that has each category of character (e.g., success/failure, survived/died, infected/not infected, male/female).

We have also talked about using the Central Limit Theorem to model the sampling distribution of the mean of a character as a normal distribution. This works for any kind of data *so long as the sample size is big enough*. Knowing when the sample size is big enough can be a tough ask, and many times we can't get any more data. What do we do when we don't have enough data?

It turns out that we know the exact sampling distribution for many estimates from data that are normally distributed.

For now, let's assume we have data that are normally distributed. This means that the individual character we might measure, call it *X*, can be modeled as follows

$$X \sim \text{Normal}(\underset{\underset{\text{mean}}{\uparrow}}{\mu}, \underset{\underset{\text{variance}}{\nwarrow}}{\sigma^2})$$

Typical examples include individual height, weight, size, productivity, concentrations, pH, and many others. The normal distribution shows up frequently (but not always!) so it's important to know about.

What the distribution of *X* tells us is that, in the population of interest, there is a mean and a variance such that the distribution of individual characters in the population follows something close to a "bell curve".

Scientists are often interested in estimating either the mean or the variance. Let's start with the mean first.

Estimating the mean from a normal population

If our interest is in the mean of the population, we know we can use the sample mean to estimate it. The sample mean is denoted by the symbol “*X-bar”* and is calculated using the equation

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

individual data points

Sample mean

sample size

What is the sampling distribution of this sample estimate? We learned before that we could use the Central Limit Theorem to write the sampling distribution as a normal. But this is only an approximation (an approximation that works really well when you have a large sample size).

But if each *Xi* are normally distributed and the sample is a random sample, then the sampling distribution of the sample mean can be determined with a ***t*-distribution.**

The *t*-distribution

The *t*-distribution doesn’t describe the mean. Instead, it measures the distance of the sample mean to the population mean in units of standard error. Here is the equation to calculate a t-statistic that goes into the t-distribution.

$$t = \frac{\bar{X} - \mu}{SE_{\bar{X}}}$$

how far sample mean, $\bar{X}$, is from population mean, μ.

standardized by the standard error

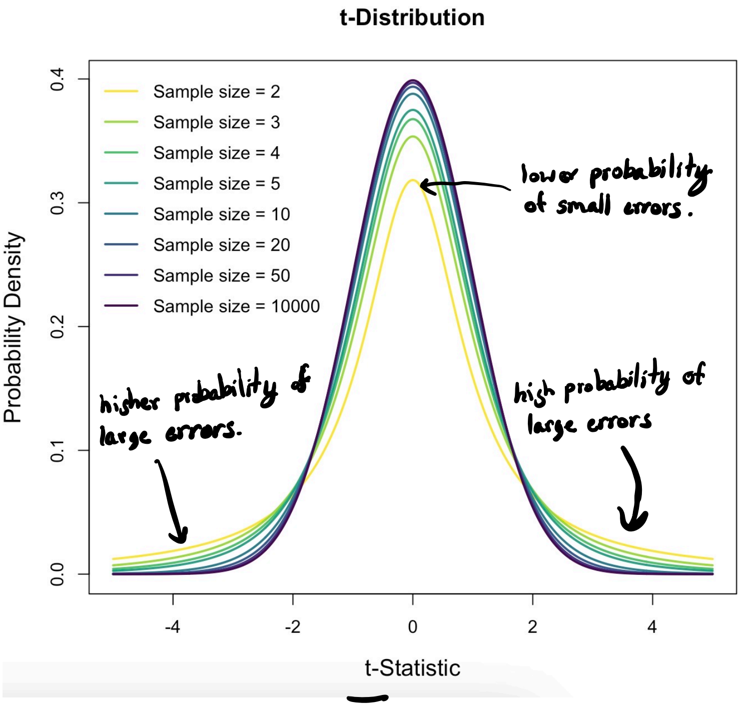
T-values are just like z-scored. When the value is 1, it means a mean calculated from a mean is 1 standard error (i.e., a standard deviation of the sampling distribution) above the population mean. When it is -2, this signifies a case where the mean calculated from a sample is 2 standard errors below the true value. Hence, *t* values indicate how wrong a sample mean is in standardized units.

The *t*-distribution shows all possible *t*-values (from -infinity to +infinity) and their associated probabilities. Thus, it tells us the probability distribution for how wrong the estimate of the mean from a sample is.

Just like the binomial, Bernoulli, and the normal distribution before, the *t*-distribution has parameters. The binomial has two parameters: the sample size, *n*, and the probability of success, *p*.

The *t*-distribution has a single parameter called **the degrees of freedom, which is the sample size - 1.** We typically write degrees of freedom as "*df*" and we typically write the sample size as *n* (for number of individuals in the sample). As such, *df* = *n* -1.

Here is the distribution of the *t*-statistic for different sample sizes.



lower probability of small errors.

higher probability of large errors.

high probability of large errors

$$t = \frac{\bar{X} - \mu}{SE_{\bar{X}}} \leftarrow$$ Standard measure of error in estimate.

What you can see from this is that the *t*-distribution looks a bit like a normal distribution. As a reference, the dark purple line is pretty darn close to a normal distribution.

But the *t*-distribution differs in an important way. The spread of the *t*-distribution is much larger than the spread of the normal distribution. This represents the fact that the *t*-distribution can account for larger errors associated with small sample sizes. Especially when sample sizes are small (purple and green lines), there is a much better chance of estimating a mean from a sample far away from the true mean.

Luckily, what counts as small isn't too small. You can see the distribution barely distinguishes samples with 50 individuals from samples with 10,000 individuals. Here, the *t*-distribution looks nearly identical to a normal distribution, which we would use under the Central Limit Theorem.

How do we use this to make confidence intervals? We use the cumulative distribution function to find quantiles of interest. Image we want a 95% confidence interval. In that case, we need the 2.5% and 97.5% quantiles. We can find them using R like this.

$t_{0.025,df} = qt(0.025,\ df = n-1) \longleftarrow$ 2.5% quantile

$t_{0.975,df} = qt(0.975,\ df = n-1) \longleftarrow$ 97.5% quantile

These both mean

$Pr(t < t_{0.025,df}) = 0.025$ and

$Pr(t < t_{0.975,df}) = 0.975$

Thus $Pr(t_{0.025,df} < t < t_{0.975,df}) = 0.95.$

Because $t = \frac{\bar{x} - \mu}{SE_{\bar{x}}}$, we have

$$Pr\left(t_{0.025,df} < \frac{\bar{x} - \mu}{SE_{\bar{x}}} < t_{0.975,df}\right) = 0.95.$$

Rearranging inside the probability gives

$$Pr\left(\underset{\text{lower value of confidence interval.}}{\underline{\bar{x} - t_{0.975,df} SE_{\bar{x}}}} < \mu < \underset{\text{upper value of Confidence Interval.}}{\underline{\bar{x} - t_{0.025,df} SE_{\bar{x}}}}\right) = 0.95.$$

Note: Because the t-distribution is symmetric $-t_{0.025,df} = t_{0.975,df}$ and $-t_{0.975,df} = t_{0.025,df}$, which leads to more commonly written (equivalent form)

$$Pr\left(\bar{X} + t_{0.025,df} \cdot SE_{\bar{X}} < \mu < \bar{X} + t_{0.975,df} \cdot SE_{\bar{X}}\right) = 0.95$$

(Same as above)

Example.

We are interested in the average CD4+ lymphocyte counts in the human population. We randomly select $n = 30$ individuals and measure their CD4+ counts. From this sample, the mean and standard deviation are

$$\bar{X} = 700 \text{ cells/mm}^3$$

$$s = 120 \text{ cells/mm}^3$$

What is the 95% confidence interval of the mean?

$$SE_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{120 \text{ cells/mm}^3}{\sqrt{30}} \qquad df = n-1 = 30-1 = 29$$

From above, our confidence interval is

$$\left[\bar{X} + t_{0.025,df} \cdot SE_{\bar{X}} \, , \, \bar{X} + t_{0.975,df} \cdot SE_{\bar{X}}\right]$$

$$\left[700 + t_{0.025,29} \cdot \frac{120}{\sqrt{30}} \, , \, 700 + t_{0.975,29} \cdot \frac{120}{\sqrt{30}}\right]$$

$$t_{0.025,29} = qt(0.025, df=29) = -2.045 \nwarrow \text{From R}$$

$$t_{0.975,29} = qt(0.975, df=29) = 2.045 \swarrow$$

$$\left[700 - 2.045 \times \frac{120}{\sqrt{30}}, 700 + 2.045 \times \frac{120}{\sqrt{30}}\right]$$

$$[655.19, 744.81].$$

This interval is slightly larger than if we had used a normal distribution. With a normal distribution, the corresponding quantiles (z-scores) are -1.96 and 1.96 for the 2.5% and 97.5% quantiles, respectively. In that case, the 95% confidence interval is [657.06, 742.94].

The t-distribution here makes slightly larger confidence intervals to account for the fact that the sample size is 30, rather than “very many”. If instead we had 10 individuals in our sample, the t quantiles are -2.26 and 2.26 giving a confidence interval of [650.44, 749.56].

Remember, we can do all the same for different alpha levels of the confidence interval for finding the correct quantiles. All else is the same.

Estimating the variance from a normal distribution

Imagine that we are not interested in the mean, but rather the variance. For example, I might want to estimate how much variability there is in some phenotype because phenotypic variability is required for natural selection. Homogeneous populations don’t evolve by natural selection (they can evolve for other reasons, such as mutation).

To do that, estimating the mean doesn’t make much sense. We estimate the variance from samples with the *sample variance*:

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

You might notice that the sample variance adds us a bunch of terms. Indeed, the sample variance adds up squared deviations for each observation in a sample of data. As such, we could use the Central Limit Theorem because it applies whenever we add up very many identical random variables. But it turns out you need quite a few to make this work.

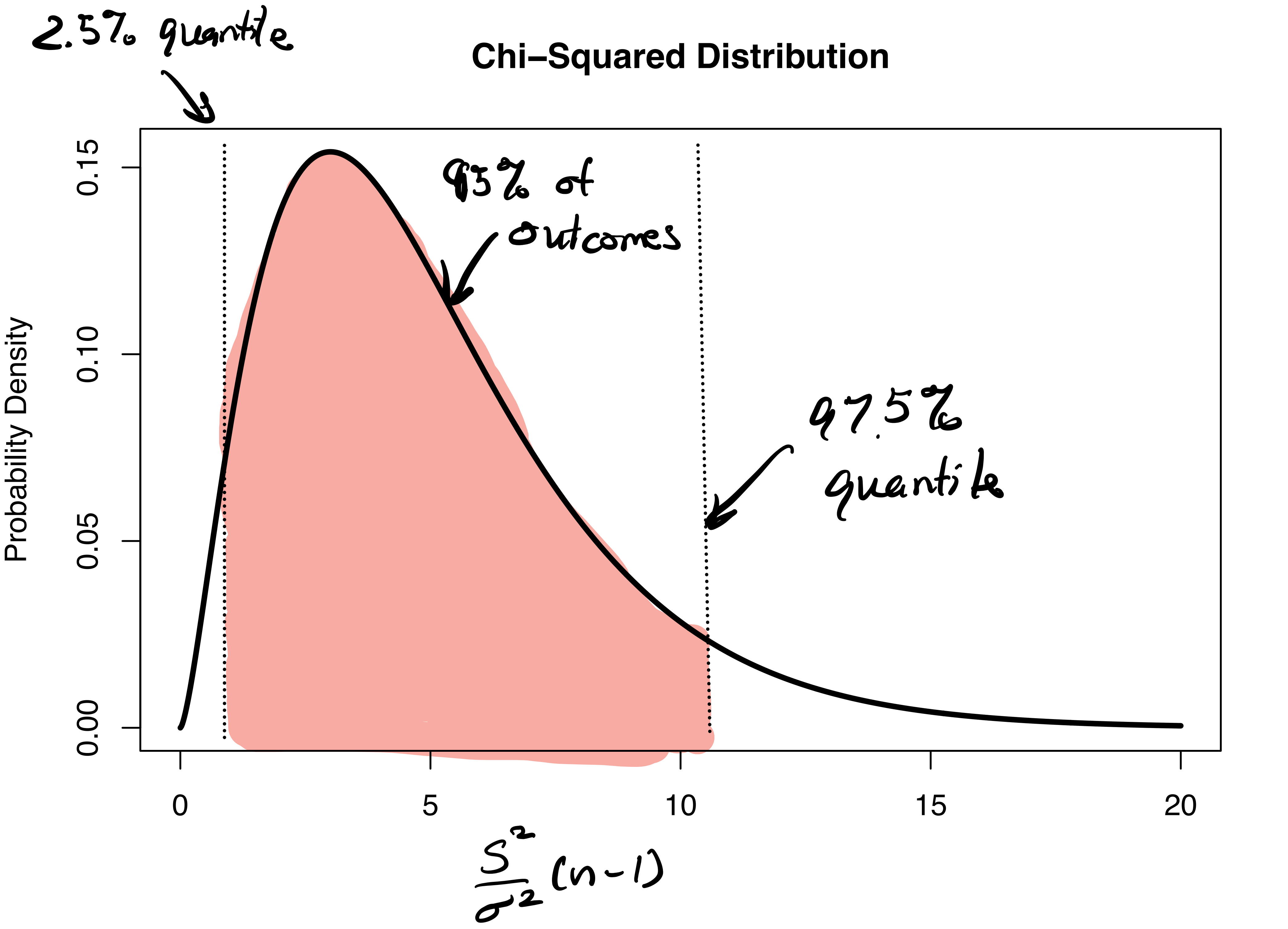
Luckily, for normally distributed data, we have a sampling distribution that helps us make confidence intervals for the variance of a population. The sampling distribution, much like the t-distribution, doesn't directly model the sample variance, but gives the probability for *how close the sample variance is to the population variance.* In exact terms, we have a sampling distribution for

$$\frac{\overset{\xleftarrow{\text{sample variance}}}{S^2}}{\underset{\xleftarrow{\text{population variance}}}{\sigma^2}}(n-1) \xleftarrow{\text{degrees of freedom}}$$

Notice that *S*-squared is capitalized in this equation. This represents the fact that *S-squared* is a random variable and can take many values with different probabilities. The crucial part of this equation is the ratio of the sample to the population variance. When this ratio is 1, the sample variance matches the population variance. When the ratio is less than 1, the sample variance underestimates the population variance. When the ratio is greater than 1, the sample variance overestimates the population variance.

The sampling distribution for this ratio is called a "Chi-Squared" distribution. Chi-squared distributions always take values great than 0 and like the *t*-distribution, chi-squared distributions have a single parameter, the degrees of freedom. The degrees of freedom is equal to one less than the sample size, i.e., *n* - 1. Here it written out in math.

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


2.5% quantile
Chi–Squared Distribution
95% of outcomes
97.5% quantile
Probability Density
0.00
0.05
0.10
0.15
0
5
10
15
20
$\frac{S^2}{\sigma^2}(n-1)$

We can find quantiles from this just like any other distribution. The functions to use in R are

- dchisq - find probability values
- pchisq - find cumulative probability values
- qchisq - find quantiles
- rchisq - pull out random values

We will use qchisq() to find quantiles that give 95% of the most likely outcomes.

$$Pr\left(\frac{S^2}{\sigma^2}(n-1) < \chi^2_{0.025,df}\right) = 0.025$$

↖ 2.5% quantile

$$\chi^2_{0.025,df} = qchisq(0.025,\ df = n-1)$$

$$Pr\left(\frac{S^2}{\sigma^2}(n-1) < \chi^2_{0.975,df}\right) = 0.975$$

↖ 97.5% quantile

$$\chi^2_{0.975,df} = qchisq(0.975,\ df = n-1)$$

$$Pr\left(\chi^2_{0.025,df} < \frac{S^2}{\sigma^2}(n-1) < \chi^2_{0.975,df}\right) = 0.95$$

We can then rearrange this equation in parenthesis to get

$$Pr\left(\frac{S^2(n-1)}{\chi^2_{0.975,df}} < \sigma^2 < \frac{s^2(n-1)}{\chi^2_{0.025,df}}\right) = 0.95$$

This is a 95% confidence interval for the population variance. That is, it gives the range of values where the population variance lies with probability of 95%. S^2 is calculated from a sample and *n* is defined by the sample. All that is left to calculate are the specific quantiles you might want based on the degrees of freedom (again, *df* = *n* -1).

Example

In the CD4+ example above, we have all we need.

$s^2 = (120 \text{ cells/mm}^3)^2 = 14{,}400 \text{ cells}^2/\text{mm}^6.$ ← units have to be squared!

$n = 30$

$df = n - 1 = 29.$

$\chi^2_{0.025,29} = qchisq(0.025, df = 29) = 16.05$

$\chi^2_{0.975,29} = qchisq(0.975, df = 29) = 45.72$

$$\frac{S^2(n-1)}{\chi^2_{0.975,29}} < \sigma^2 < \frac{S^2(n-1)}{\chi^2_{0.025,29}}$$

$$\frac{14{,}400 \times 29}{45.72} < \sigma^2 < \frac{14{,}400 \times 29}{16.05}$$

$$9133 < \sigma^2 < 26{,}023$$ ← 95% confidence interval.

If we want this in standard deviations, we just take the square root.

95% confidence interval of σ

$$\sqrt{9133} < \sqrt{\sigma^2} < \sqrt{26{,}023}$$

$$95.57 \text{ cells}/\text{mm}^3 < \sigma < 161.32 \text{ cells}/\text{mm}^3$$

If instead we had $n=10$ individuals in our sample, then the 95% confidence interval is

$$82.54 \frac{\text{cells}}{\text{mm}^3} < \sigma < 219.07 \text{ cells}/\text{mm}^3$$