You have studied bill width in a population of finches for many years. You record your data in units of the standard deviation of the population, and you subtract the average bill width from all of your previous studies of the population. Thus, if the bill widths are not changing year-to-year and they are distributed according the Normal distribution (as many quantitative traits are), then your data should be described by $\mathcal{N}(0, 1)$.

Problem: Consider the following data set collected from 5 randomly sampled birds following from this population, but following a year of drought:

Indiv.	standardized bill width
1	0.01009121
2	3.63415088
3	-1.40851589
4	3.70573177
5	-0.94145782

Oddly enough it appears that you have enough money to measure bill widths using an SEM (based on the ridiculous number of digits in your measurments), but you can only catch 5 finches.
Can you conclude that the mean bill width in the population has changed?

Solution: If you talk to someone without any statistical training (and then translate their answer into stats jargon), they will say that we should answer this by:

1. estimating the mean based on the data (estimate $\hat{\mu}$), and then
2. see if $\hat{\mu} = 0$.

But this clearly ignores the fact that our estimate will be affected by sampling error (so we'll conclude there has been a change in the mean any time we do the test).

If you ask someone trained in Neyman-Pearson style of hypothesis testing they'll say that we should:

1. state the *null hypothesis* ($H_0 : \mu = 0$, in this case);
2. stat the *alternative hypotheses* ($H_A : \mu \neq 0$ in this case);
3. choose a *test statistic*;
4. choose your *Type I error rate* (usually denoted α) to be what you consider to be an acceptable probability of rejecting a null hypothesis when it is true.
5. determine the *null distribution* of the test statistic - the frequency distribution of the values of the statistic that we would see if the null hypothesis were true.
6. from the null distribution of the test statistic, the Type I error rate, and the knowledge of what values of the test statistic are more compatible with the alternative hypothesis than the null, you can determine the *critical value* of your test statistic.
7. If the value of the test statistic calculated on the real data is more extreme than the critical value, then you reject the null.

This is a general procedure. Which test statistic you should use is not always obvious.

The Neyman-Pearson lemma states that we should use the likelihood ratio test-statistic if we are testing two distinct points (e.g. if $H_0 : \mu = 0$ and $H_A : \mu = 1$ for example). The lemma actually states that the likelihood ratio test-statistic is the most powerful (e.g. gives us the most power to reject the null when it is false) that is "honest" in the sense of guaranteeing it reported Type I error rate.

In our case, we do not have a distinct hypothesis ("point hypothesis") as an alternative. But we can still use the likelihood ratio as our test statistic, but we have to have a pair of points to use when calculating the test statistic. We can use the MLE when calculating the likelihood that is in the denominator in the ratio, and we can use the null hypothesis' value of μ when calculating the numerator. We also take the log of the ratio and multiply it by -2 (so that the properties that we proved in problem #3 of homework #2 hold):

$$\begin{aligned} \Lambda &= L(\hat{\mu})/L(\mu_0) \\ D = -2\ln\Lambda &= 2\ln[L(\hat{\mu})] - 2\ln[L(\mu_0)] \end{aligned}$$

For a dataset that is series of independent draws from the same normal:

$$\begin{aligned} p(X \mid \mu, \sigma) &= p(X|\mu,\sigma) \\ &= \prod_{i=1}^{n} p(x_i|\mu,\sigma) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ \ln[p(X \mid \mu, \sigma)] &= \sum_{i=1}^{n} \ln\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right] \\ &= \sum_{i=1}^{n} \left(\ln\left[(2\pi)^{-1/2}\right] + \ln\left[\sigma^{-1}\right] + \ln\left[e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right]\right) \\ &= n\ln\left[(2\pi)^{-1/2}\right] + n\ln\left[\sigma^{-1}\right] + \sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{2\sigma^2} \\ &= C_1 + n\ln\left[\sigma^{-1}\right] + \sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{2\sigma^2} \end{aligned}$$

If we know σ and are only interested in estimating μ, then collapsing the constant-with-respect-to-μ terms yields:

$$\begin{aligned} \ln[p(X \mid \mu, \sigma)] &= C_2 + \sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{2\sigma^2} \\ &= C_2 - \sum_{i=1}^{n} \frac{x_i^2 - 2\mu x_i + \mu^2}{2\sigma^2} \\ &= C_2 - \sum_{i=1}^{n} \frac{x_i^2}{2\sigma^2} - (-2\sigma^2)^{-1} \sum_{i=1}^{n} \left(2\mu x_i + \mu^2\right) \\ &= C_3 - (2\sigma^2)^{-1} \sum_{i=1}^{n} \left(-2\mu x_i + \mu^2\right) \\ &= C_3 - (2\sigma^2)^{-1}(-2\mu n\bar{x} + n\mu^2) \\ &= C_3 - \frac{n(-2\mu\bar{x} + \mu^2)}{2\sigma^2} \end{aligned}$$

We can find, unsurprisingly that the MLE is $\hat{\mu} = \bar{x}$:

$$\frac{\partial \ln[p(X \mid \mu, \sigma)]}{\partial \mu} = -\frac{n(-2\bar{x} + 2\mu)}{2\sigma^2}$$

$$\begin{aligned}\frac{-n}{\sigma^2}(\hat{\mu}-\bar{x}) &= 0\\ \hat{\mu} &= \bar{x}\end{aligned}$$

The log likelihood at the MLE is:

$$\begin{aligned}\ln[p(X \mid \mu=\bar{x},\sigma)] &= C_3 - \frac{n(-2\bar{x}^2\bar{x}^2)}{2\sigma^2}\\ &= C_3 + \frac{n\bar{x}^2}{2\sigma^2}\end{aligned}$$

We may recall from biostats/biometry that the cutoff for a plausible values for μ (the ends of the 95%CI for μ) are given by $\bar{x} \pm \frac{1.96\sigma}{\sqrt{n}}$. Where the 1.96 is the critical value from the Z (standard normal) distribution. If we were going base a confidence interval or a hypothesis test on the LRT, then we might want it to go from significant (at $\alpha = 0.05$) at this point. So what is the log-likelihood for $\bar{x} + \frac{1.96\sigma}{\sqrt{n}}$?

$$\begin{aligned}\ln\left[p\left(X \mid \mu=\bar{x}+\frac{1.96\sigma}{\sqrt{n}},\sigma\right)\right] &= C_3 - \frac{n\left(-2\left(\bar{x}+\frac{1.96\sigma}{\sqrt{n}}\right)\bar{x}+\left(\bar{x}+\frac{1.96\sigma}{\sqrt{n}}\right)^2\right)}{2\sigma^2}\\ &= C_3 - \frac{n\left(-2\bar{x}^2-\frac{2\bar{x}1.96\sigma}{\sqrt{n}}+\bar{x}^2+\frac{2\bar{x}1.96\sigma}{\sqrt{n}}+\left[\frac{1.96\sigma}{\sqrt{n}}\right]^2\right)}{2\sigma^2}\\ &= C_3 - \frac{n\left(-\bar{x}^2+\frac{3.84\sigma^2}{n}\right)}{2\sigma^2}\\ &= C_3 + \frac{n\bar{x}^2}{2\sigma^2} - \frac{3.84}{2}\end{aligned}$$

So the critical value for the LRT is:

$$\begin{aligned}D_{\alpha=0.05} &= 2\ln[L(\hat{\mu})] - 2\ln[L(\mu_0)]\\ &= 2\left[C_3+\frac{n\bar{x}^2}{2\sigma^2}\right] - 2\left[C_3+\frac{n\bar{x}^2}{2\sigma^2}-\frac{3.84}{2}\right]\\ &= 3.84\end{aligned}$$

So we would reject any null for μ that results in a log likelihood that is 3.84 units worse than the log likelihood at the MLE.

You may recall that $\chi^2_{\alpha=0.05,df=1} = 3.84$. This similarity in the number is not a coincidence.

The null distribution of the LR test statistic in cases like this one (in which the MLE is not at a boundary,

If we test 2 models and:

- One model is more general. And the "smaller" model is nested inside the more general model; and
- The smaller model can be obtained by making k constraints on the parameters of the more general model to values. Note: these constraints (if they are constraints to specific values) that are to values that are "inside" their legal ranges (*i.e.* not parameter values that coincide with the boundary of the parameter's legal range); and

- your dataset is not too tiny

then: The null distribution of the LRT approximately simply χ^2_k where the degrees of freedom, k, is simply to the number of parameters that constrained to take on certain values.