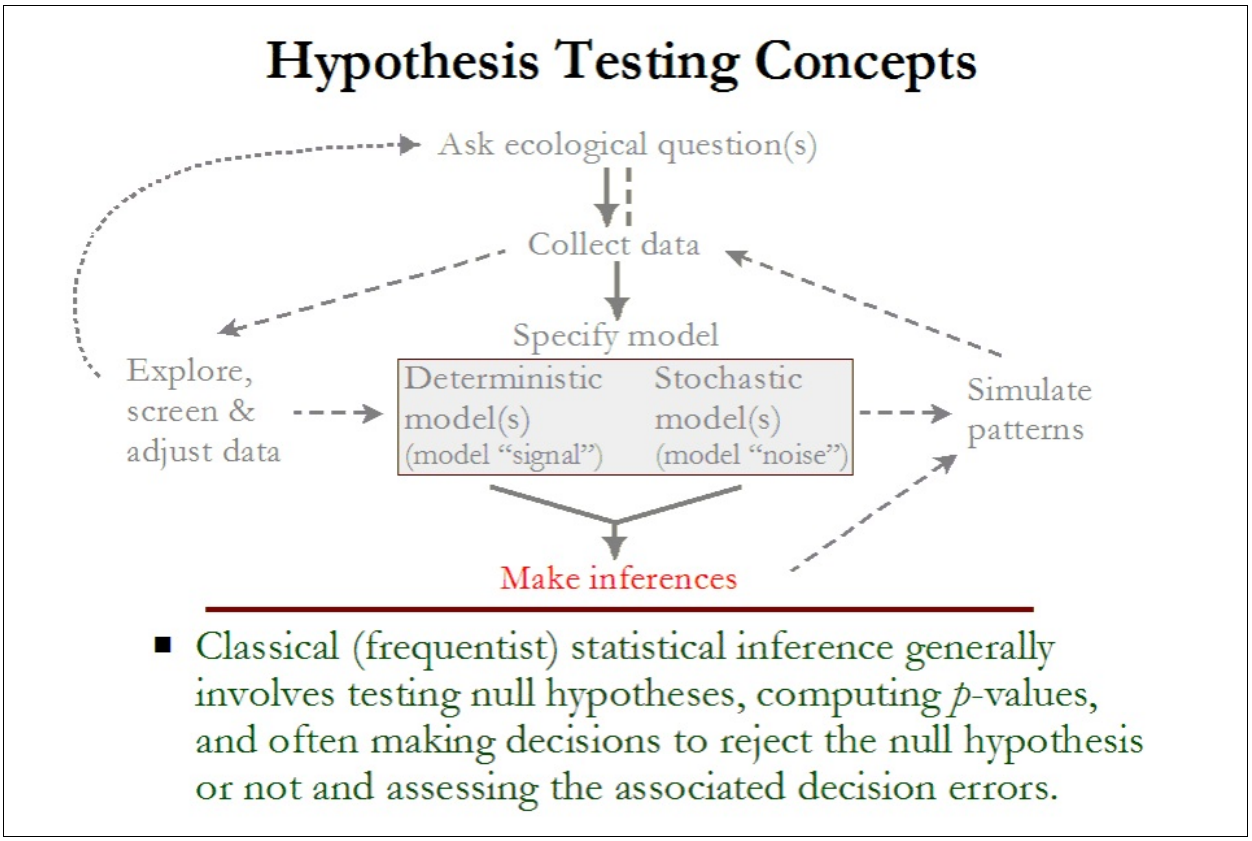
# Analysis of Environmental Data

## Chapter 6b. Conceptual Foundations:

*Hypothesis Testing Concepts*




## 1. Was is null hypothesis testing?

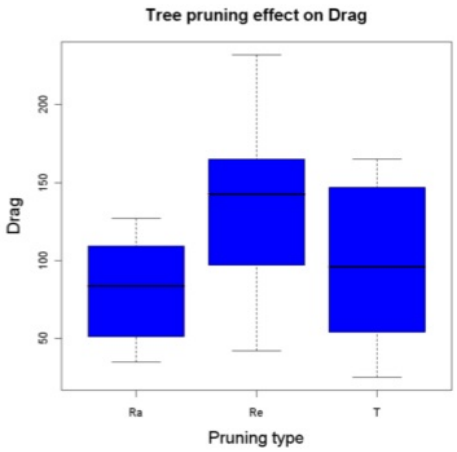
Recall that statistical inference involves confronting models with data to estimate model parameters, test hypotheses, compare alternative models and make predictions, and there are different inference frameworks for conducting such inference. Classical Frequentist inference involves heavy reliance on hypothesis testing. Indeed hypothesis testing has been the trademark of statistical inference for much of the past century, and only recently has the emphasis shifted away from hypothesis testing, to other aspects of statistical inference such as parameter estimation and model comparison. Nevertheless, because hypothesis testing is so deeply rooted in statistical inference and is still perhaps the dominant form of statistical inference, it is critical to understand some of the concepts that underpin hypothesis testing before we delve into the details of the various inference frameworks.

## Hypothesis Testing Concepts

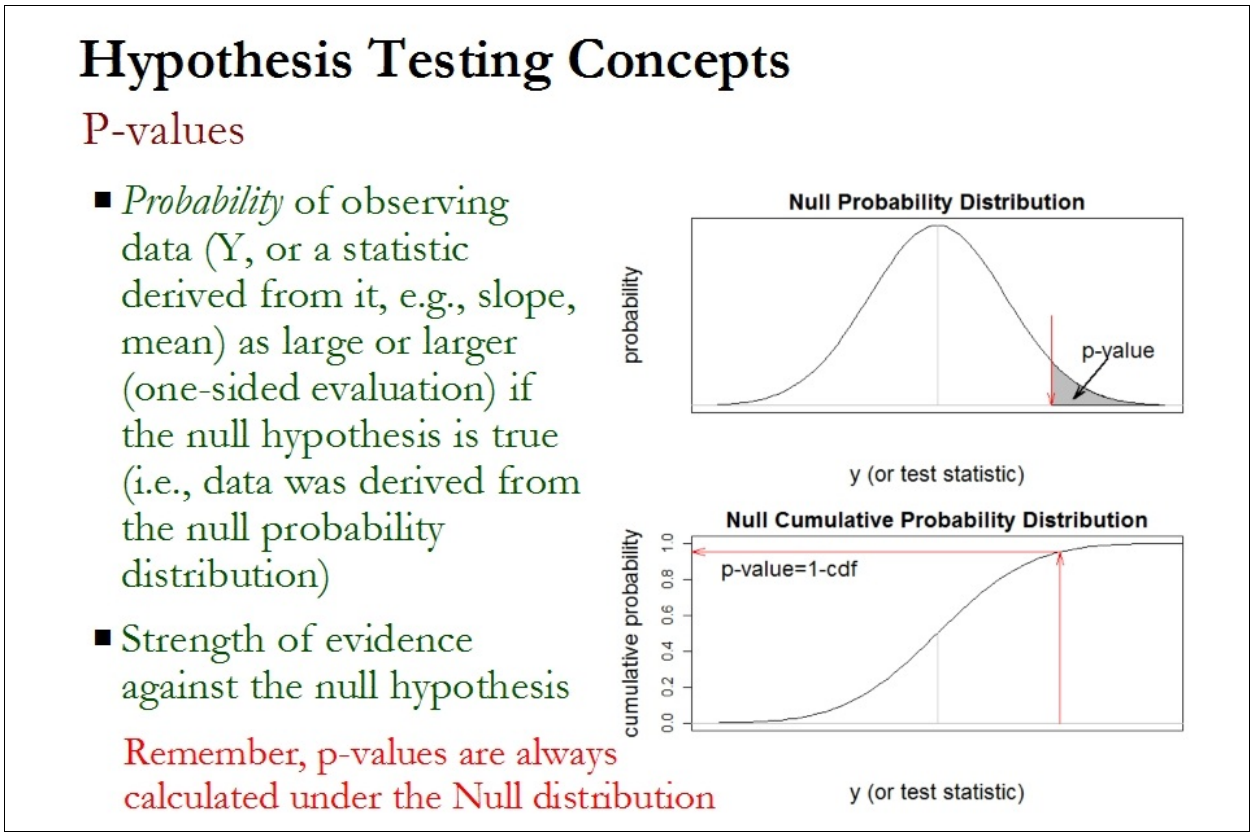
### Null hypotheses

- Statement about the system under investigation that you are trying to disprove
- Specified in terms of the statistical model, which provides a probability distribution against which to compare your data
- Usually (but not always) a statement of "no effect" or "no relationship"

What is a reasonable null hypothesis?




Hypothesis testing generally revolves around the *null hypothesis*, which is a statement about the system under investigation that you are trying to disprove. The null hypothesis is always specified in terms of the proposed statistical model, which provides a probability framework against which to compare your data. Usually the null hypothesis is a statement of "no effect" or "no relationship", which often translates into a model parameter equal to zero. For example, in the brown creeper example, a reasonable null hypothesis is that there is no relationship between brown creeper abundance and the extent of late-successional forest, which would translate into a slope of zero in the model. Similarly, in the tree pruning example, a reasonable null hypothesis is that drag does not differ among pruning types, which would translate into a single mean across pruning types. However, it is important to recognize that the null hypothesis does not have to represent the absence of a relationship. There are many circumstances where that would be a "silly" null hypothesis that is virtually guaranteed to be falsified by any dataset. For example, if we were interested in the relationship between stream discharge and watershed area, a null hypothesis of no relationship is silly since we are certain to reject it with any reasonable dataset. A more interesting null hypothesis would be that the relationship is linear; i.e., that stream discharge increases by a constant amount for every unit increase in watershed area. Deviation from this expectation then might reveal something interesting about the hydrological properties of the watersheds under investigation.




## 2. *P*-values

The null hypothesis provides a probability framework against which to compare our data. Specifically, through the proposed statistical model, the null hypothesis can be represented by a probability distribution, which gives the probability of all possible outcomes if the null hypothesis is true; it is a probabilistic representation of our expectations under the null hypothesis.

We can specify the null probability distribution for our data (random variable y) or for any statistic derived from our data (e.g., the mean, or the slope). Recall that there are probability distributions for random variables (e.g., normal, gamma, binomial, poisson, etc.) and there are probability distributions for test statistics derived from our data (e.g., *t*, *F*, chi-square). For now, it doesn't matter whether we are evaluating a random variable or a test statistic, the concepts described below are the same. In addition, it doesn't matter what probability distribution we are working with, the concepts described below are the same. But for purposes of familiarity, let's let our probability distribution represent a normally distributed random variable.

In this context, the null probability distribution (top figure) represents the probability of observing any particular value of our random variable *y* if the null hypothesis is true. The probability of observing a value of *y* as large or larger than the one we observed (i.e., for a one-sided evaluation, see below) under the null distribution (i.e., if the null hypothesis is true) is known as the *p-value*, which was originally proposed by the famous early statistician Sir Ronald Fisher. The *p*-value for this

one-sided evaluation is equal to the proportion of the null distribution that is to the right of our observed value. According to Fisher, the *p*-value can be viewed as a measure of the strength of evidence against the null hypothesis. The smaller the *p*-value the less likely it is that we would have observed that particular value if it came from the null distribution, and thus the greater the evidence that the null hypothesis is wrong.

Since the *p*-value is computed (for a one-sided evaluation) as the proportion of the null distribution that is as larger or larger than the observed value (i.e., the tail of the probability distribution), we can easily calculate the *p*-value from the cumulative probability distribution. Recall that the cumulative probability distribution gives the probability of being less than or equal to any particular value, so the *p*-value must be its compliment (1-cdf(y)). Note, if we are computing the p-value for a discrete probability distribution (e.g., Binomial, Poisson, Geometric, etc.), then we would be interested in finding the compliment of the cdf(y) for the discrete outcome that is one less than the value of interested. For example, if we were interested in the probability of observing 8 successes of out of 10 trials given a per trial probability of success = 0.3, as given by the binomial distribution, we would compute the cumulative probability of observing 7 or less successes and then take its compliment to get the probability of observing 8 or more successes. For a continuous probability distribution (e.g., Normal, Gamma, Exponential, etc.), we simply compute the cumulative probability of observing the observed value or less and then takes its compliment, because with a continuous probability distribution the next smallest possible outcome from the observed value is smaller by an infinitely small amount such that the cumulative probability asymptotically approaches the same value.

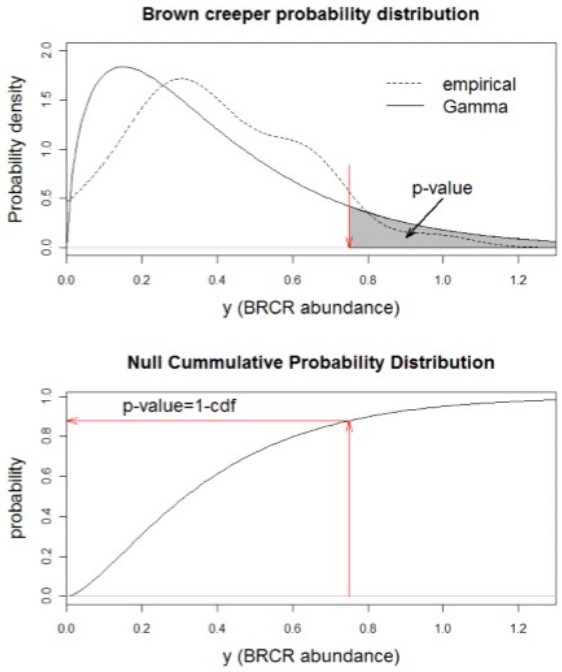
# Hypothesis Testing Concepts

## P-values

- *Random variables...* Probability of observing a value of Y as large or larger (one-sided evaluation) under the null hypothesis, for any randomly distributed variable with a sampling distribution

$H_A$: $Pr(Y_i \geq 0.75)$

Y~Gamma(shape=1.61,scale=.24)




It is important to remember that *p*-values apply equally to random variables and any statistics derived from them (since they too are "random" variables), which includes estimates of model parameters and test statistics.

For example, consider a continuously distributed random variable Y. In this example, Y is brown creeper abundance from our familiar data set, but it could be any random variable. The empirical probability distribution of Y is shown in the top figure (dashed line). Let's say we want to know the probability of observing a value of $Y_i \geq 0.75$ under the null hypothesis that Y is distributed Gamma with shape and scale parameters equal to 1.61 and 0.24, respectively. This "null" distribution is shown in the top figure (solid line). Calculating the *p*-value for $Y_i \geq 0.75$ is simple. It is the proportion of the null (Gamma) distribution to the right of 0.75, shown in the dark shaded area. Note, we could also ask what proportion of the empirical distribution is to the right of 0.75, but this represents the observed distribution and remember that we calculate *p*-values under the null distribution. We can also derive the *p*-value from the cumulative probability distribution, as it is done in practice, as shown in the bottom figure. In this case, since we are asking a one-side question; i.e., the probability of $Y_i$ greater than or equal to a particular value, we compute the *p*-value as the compliment of the cumulative probability of observing $Y_i \leq 0.75$, which in this case is roughly 0.15. Hence, we can say that there is roughly a 15% chance of observing a brown creeper abundance of greater than or equal to 0.75 if in fact the brown creeper abundance was distributed Gamma with shape=1.61 and scale=0.24.

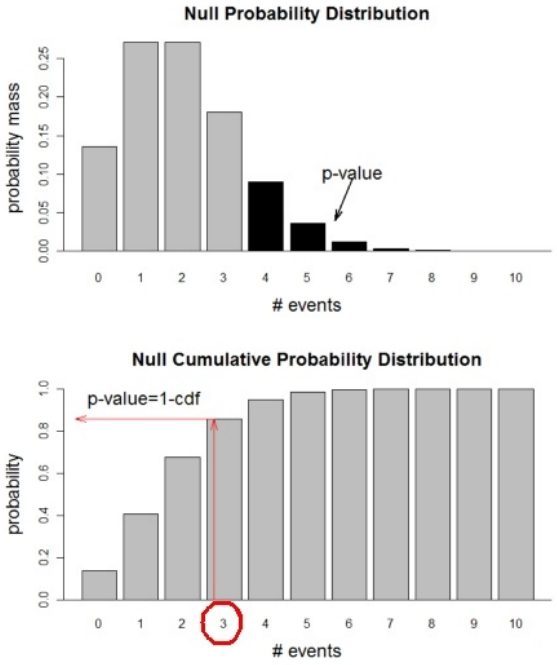
# Hypothesis Testing Concepts

## P-values

- *Random variables...*
  Probability of observing a value of Y as large or larger (one-sided evaluation) under the null hypothesis, for any randomly distributed variable with a sampling distribution

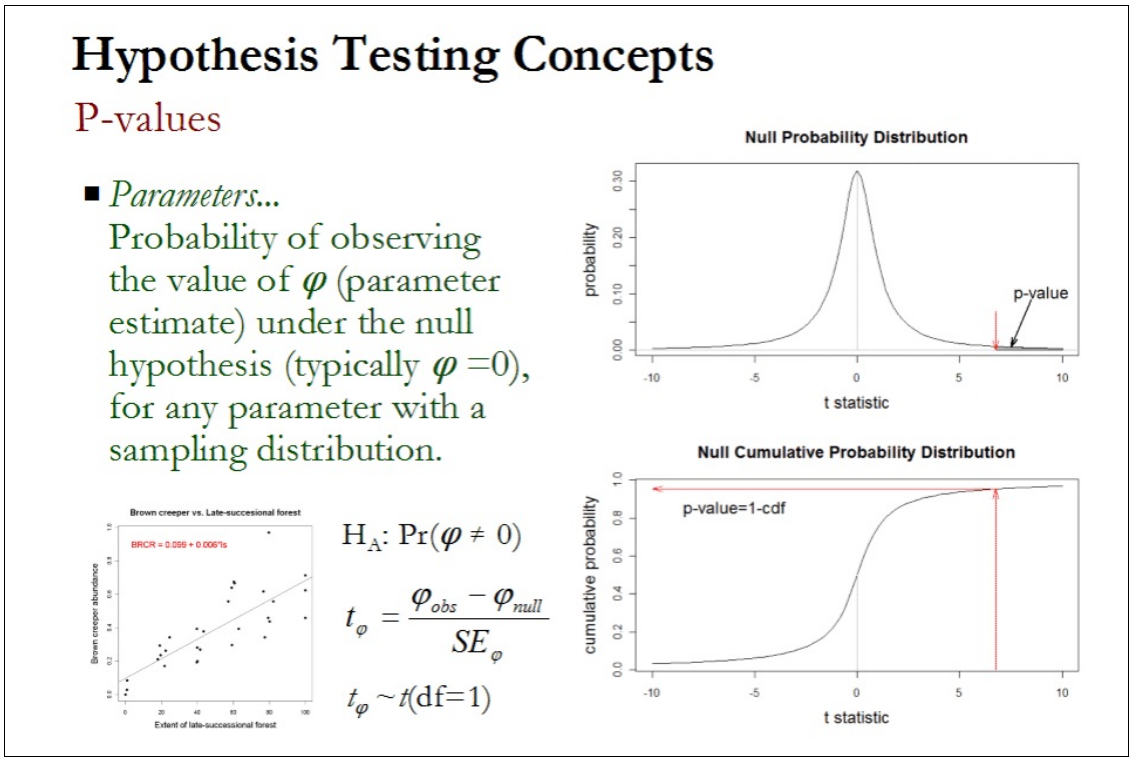
$H_A$: $Pr(Y_i \geq 4)$

Y~Poisson(*lambda*=2)

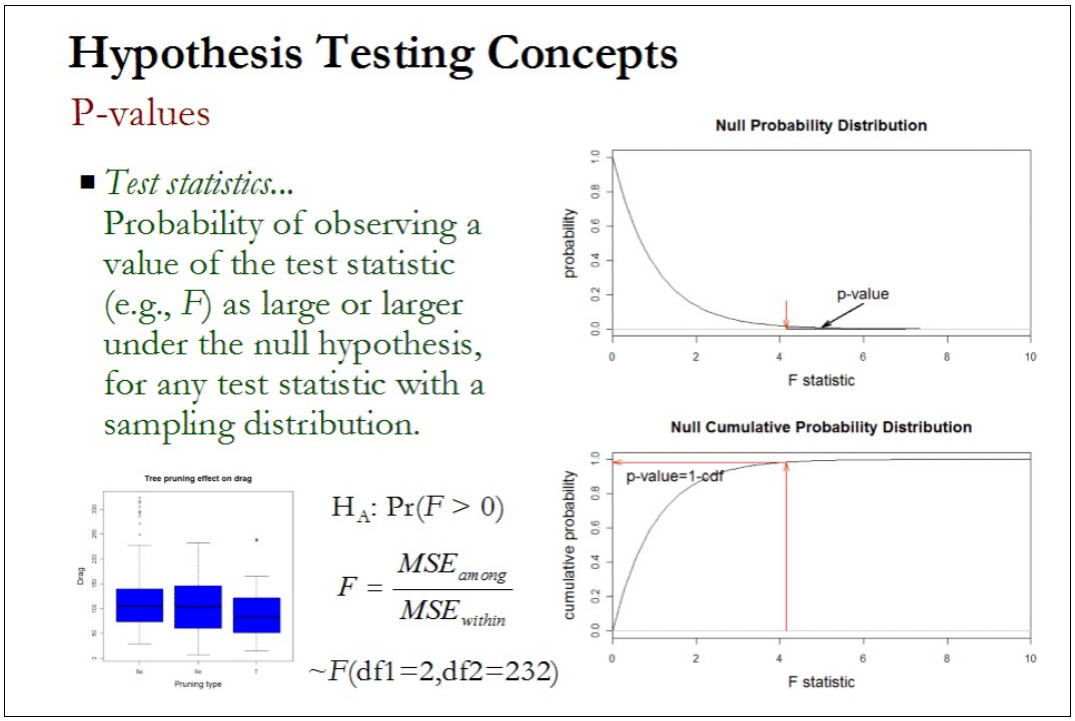



**Remember**, *p*-values apply equally to random variables with any probability distribution.

For example, consider a discretely distributed random variable Y. In this example, Y is a count of earthworms in a fixed-area plot. Let's say we want to know the probability of observing a value of $Y_i \geq 4$ under the null hypothesis that Y is distributed Poisson with Lambda (mean count) = 2. This "null" distribution is shown in the top figure. Note, it is shown as a barplot because it is a discrete distribution. Calculating the *p*-value for $Y_i \geq 4$ is simple. It is the proportion of the null (Poisson) distribution to the right of 3, shown in the dark shaded area. We can also derive the *p*-value from the cumulative probability distribution, as it is done in practice, as shown in the bottom figure. In this case, since we are asking a one-side question; i.e., the probability of $Y_i$ greater than or equal to a particular value, we compute the *p*-value as the compliment of the cumulative probability of observing $Y_i < 4$, which in this case is the same as finding the cumulative probability for the value of 3, since this represents 3 or less, and taking its compliment, which in this case results in a *p*-value of roughly 0.15. Hence, we can say that there is roughly a 15% chance of observing 4 or more earthworms in a plot if in fact the mean count of earthworms is 2 across all plots.




**Remember**, *p*-values apply equally to statistics derived from random variables. For example, a consider a model parameter φ (phi). Let's say that φ is the slope of the linear regression of extent of late-successional forest on brown creeper abundance from our familiar data set, but it could be any model parameter. Let's say that we estimated the slope using some estimator such as ordinary least squares or maximum likelihood. Now we want to know if the observed slope is statistically significant; i.e., whether it differs from zero. Because the slope is a statistic and not raw data, we can't use one of the probability distributions derived for random variables (e.g., Normal, Gamma, etc.). Instead, we need to use one of the probability distributions designed for statistics. In this case, the slope parameter can be converted into a "*t*-statistic" by taking the observed value (in this case, $b$=0.006) and subtracting the expected value under the null hypothesis ($b$=0) and dividing by the standard error of that statistic (in this case, $SE_b$=0.0008). The standard error is the standard deviation of the sampling distribution of the test statistic. We will not concern ourselves right now with exactly what a standard error is or how it is calculated. For now, suffice it to say that it "standardizes" the observed value of the statistic much the same we previously *z*-score standardized raw variables to put them on a standard scale. Once we convert the slope parameter into a *t*-statistic, we can evaluate the observed value *t* against the expected distribution under the null hypothesis of zero slope. The *t*-distribution under null hypothesis is shown in the top figure and the observed value of *t* is indicated by the red arrow. The *p*-value is the area to the right of this point, shown in dark shading. As before, we can also derive the *p*-value from the cumulative probability distribution, as shown in the bottom figure. If we ask a one-sided question; i.e., the probability of *t* greater than or equal to the observed value, we compute the *p*-value as the compliment of the cumulative probability of observing our *t* under the null hypothesis, which in this case is a very small number. Hence, we can say that there is almost no chance of observing a slope of 0.006 if in fact the true slope was zero.

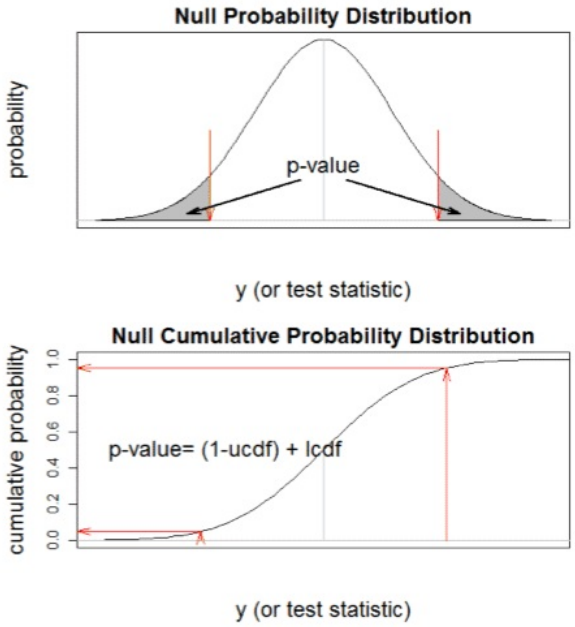



**Remember**, *p*-values apply equally to any test statistic. For example, a consider a test statistic such as *F*, which is defined for the ratio of two normally distributed variances, but it could be any test statistic. Let's take the tree pruning example and ask whether the method of pruning (3 types) has an effect on drag (i.e., wind-induced stress) on a focal tree species. We would like to know whether the variation in drag among pruning types (call this the treatment or model variance) is greater than the residual variation in drag within pruning types. In other words, does pruning type explain a significant amount of the variation in drag. This boils down to a ratio of two variances: the variance among treatment types, and the variance within treatment type. The variance among types is the variation we are trying to explain with our model, and the variation within types is the residual variation that cannot be explained by our model. The ratio of these two variances, under the assumption that they are both normally distributed, is known as an *F*-statistic (or *F*-ratio). Note, this is also known as "Analysis of Variance" or (ANOVA). Now, we want to know if the observed *F*-ratio is statistically significant; i.e., whether it differs from zero. Because the *F*-ratio is a statistic and not raw data, we can't use one of the probability distributions derived for random variables. Instead, we need to use one of the probability distributions designed for statistics. In this case, the *F*-ratio has its own probability distribution – the *F* distribution. We can compute the *p*-value for our observed *F*-ratio by evaluating the probability of observing an *F* as large or larger under the null hypothesis of no treatment effect. Under the null hypothesis, $F=0$, since none of the variance is explained by treatments (i.e., means are all equal). The *F*-distribution under null hypothesis is shown in the top figure and the observed value of *F* is indicated by the red arrow. The *p*-value is the area to the right of this point, shown in dark shading. As before, we can also derive the *p*-value from the cumulative probability distribution, as shown in the bottom figure. We compute the *p*-value as the compliment of the cumulative probability of observing our *F* under the null hypothesis, which in this case is a very small number. Hence, we can say that there is almost no chance of observing the differences among treatments if in fact the true means were the same.

# Hypothesis Testing Concepts

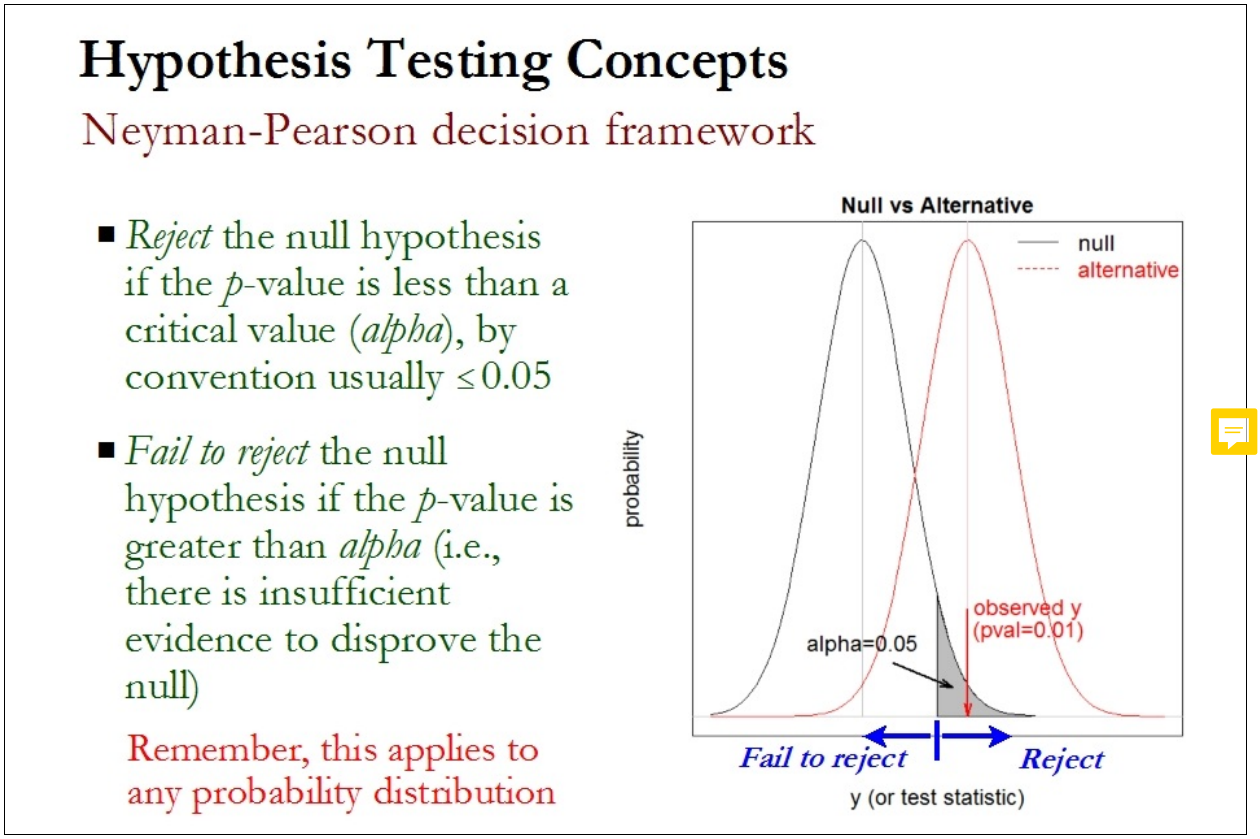
## P-values

- *One-sided test*... the probability of observing data as large or larger (or as small or smaller) if the null hypothesis is true
- *Two-sided test*... the probability of observing data as different (≤ or ≥)as observed if the null hypothesis is true



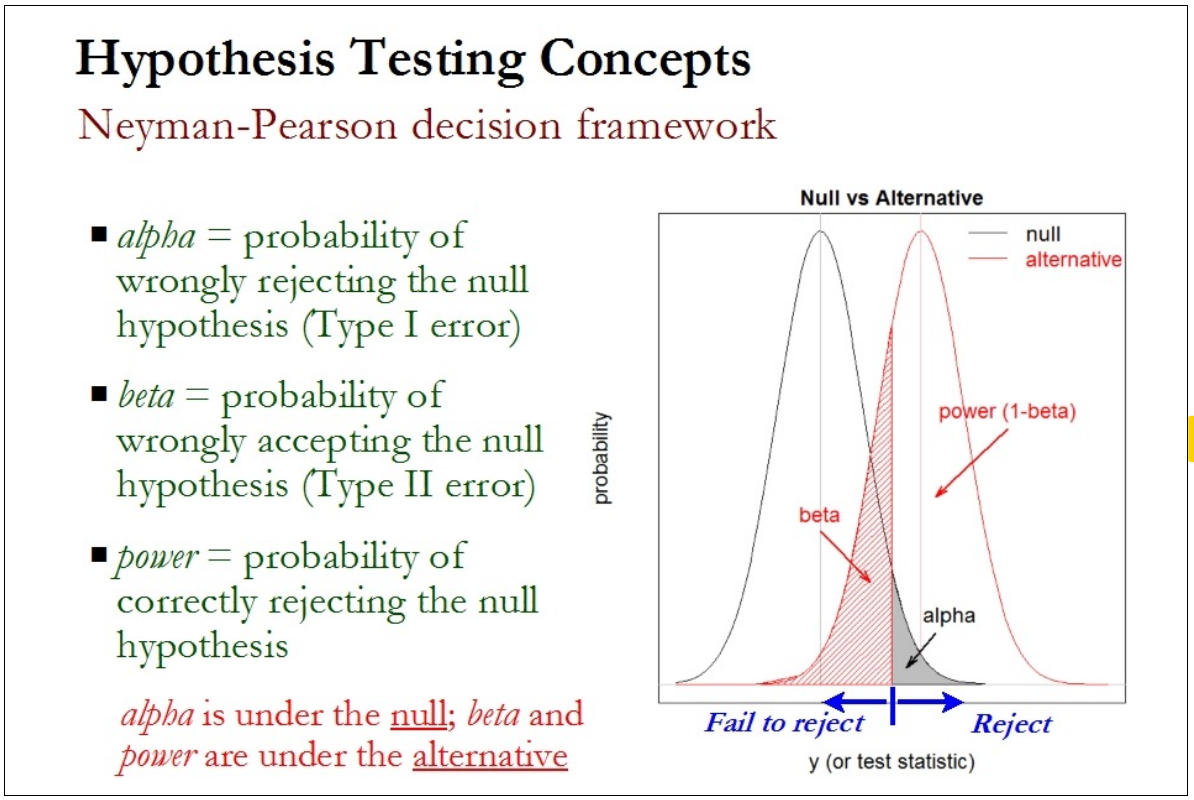

One-side test.--The interpretation given above for a *p*-value is correct for an "upper one-sided test"; i.e., when we are interested in knowing the probability of observing a value as large or larger than the one we observed if the null hypothesis is true. In this case, we are only interested in the proportion of the distribution that is to the right of our value – the upper tail of the distribution. However, we could equally be interested in knowing the probability of observing a value as small or smaller than the one we observed if the null hypothesis is true. In this case, we are only interested in the proportion of the distribution that is to the left of our value – the lower tail of the distribution. The latter would be referred to as a "lower one-sided test". The *p*-value for a lower one-sided test is computed similarly, except when computed using the cumulative distribution (as is usually done), there is no need to take the compliment since the cumulative distribution already gives the proportion of the distribution that is to the left of the observed value.

Two-sided test.–More often than not, we are not interested in testing whether the observed value is larger than expected, or conversely smaller than expected (one-sided tests), but rather whether the observed value is different (either larger or smaller) than expected. This constitutes a "two-sided test" because we are interested in both tails of the distribution. Specifically, for any particular value of Y (positive or negative) we can ask what is the probability of observing a $Y_i$ that is different from expected. If $Y_i$ is positive, we compute the upper tail probability as before (i.e., proportion to the right of the observed value) and then add it to the comparable lower tail probability (i.e., proportion to the left of a value equidistant from the expected). Using the cumulative distribution, we simply add the compliment of the upper to the lower.




### 3. Neyman-Pearson Decision Framework

Some other famous early statisticians (Jerzy Neyman and Egon Pearson) proposed that it would make sense to adopt a decision rule to decide when to "fail to reject" the null hypothesis and when to "reject" the null hypothesis in favor of an alternative hypothesis. Note, they made the notion of an alternative hypothesis explicit; i.e., that we conceptualize an alternative distribution, which is important to the concept of power discussed below. They proposed that when the *p*-value gets smaller than some critical value, known as *alpha*, that we decide to reject the null hypothesis. So, if we compute the *p*-value for our observed $Y_i$ (or test statistic) and it is smaller than *alpha*, we reject the null hypothesis in favor of our alternative hypothesis. By convention, the *alpha* level is usually set at 0.05. Thus, if the *p*-value $\leq 0.5$, we reject the null hypothesis; otherwise, we fail to reject it (technically, we can never "accept it" since we can never prove that it is true). Note, if the *p*-value is less than *alpha* and we therefore decide to reject the null hypothesis, there is a chance that we are making a mistake; i.e., that the null hypothesis is actually true and we simply observed one of the unlikely but possible outcomes. Making this mistake is known as a *Type I error*, which is simply the probability of wrongly rejecting the null hypothesis. By convention, science demands that we keep the Type I error rate very low ($\leq 0.05$), because science has deemed it unwise to accept an alternative until the evidence is overwhelmingly in favor of it – analogous to the innocent until proven guilty concept that underpins our legal philosophy.




*Beta*.–While *alpha* sets the critical *p*-value for rejecting the null hypothesis, and thus determines the Type I error rate that we are willing to accept, *beta* addresses another kind of error. *Beta* refers to the probability of wrongly accepting (or failing to reject) the null hypothesis. In other words, let's suppose the null hypothesis is actually wrong and the alternative hypothesis is correct, but we fail to reject the null hypothesis. Then we have made a mistake known as a *Type II error*, which is simply the probability of wrongly accepting (failing to reject) the null hypothesis. *Beta* is equal to the area under the alternative probability distribution corresponding to a decision to accept (fail to reject) the null hypothesis. Thus, *beta* is dictated by *alpha* because *alpha* determines the decision to accept or reject the null hypothesis. It is important to note that *beta* is associated with the alternative distribution, because *beta* refers to the case when the alternative distribution is correct (and the null is wrong). Because of the convention to control *alpha* and the Type I error rate, *beta* and the Type II error rate is an emergent property of the situation; the value of *beta* depends on how much the null and alternative distributions overlap and on the user-specified value of *alpha*.

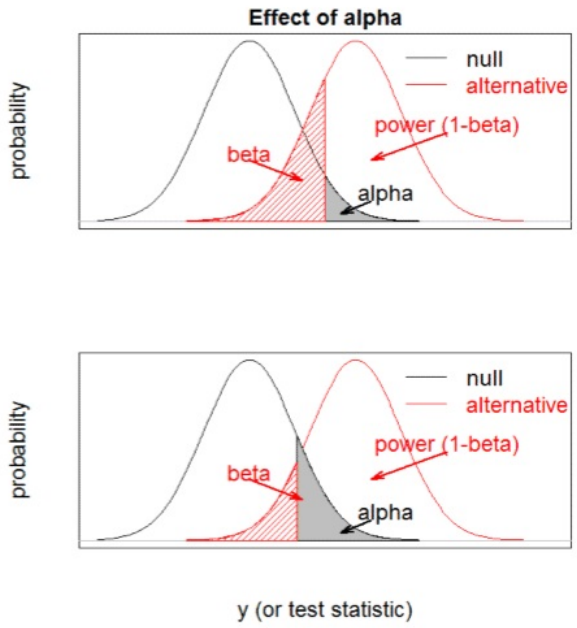
*Power*.–While *beta* gives the probability of wrongly accepting the null hypothesis, power is the probability of correctly rejecting the null hypothesis. Power is the logical compliment of *beta* under the alternative distribution. Thus, in the figure shown, power includes all of the area under the alternative distribution except that given by *beta* (i.e., it includes the area depicted as *alpha* under the null; the dark shaded area). If the alternative hypothesis is true, *beta* gives the probability of making the wrong decision (to accept the null), while power gives the compliment – the probability of making the right decision. Arguably, power is something we want lots of – "statistical" power that is! We generally want to be able to reject the null hypothesis in favor of our alternative if it is the right thing to do, so we would like to constructs tests that have high power.

# Hypothesis Testing Concepts

## Neyman-Pearson decision framework

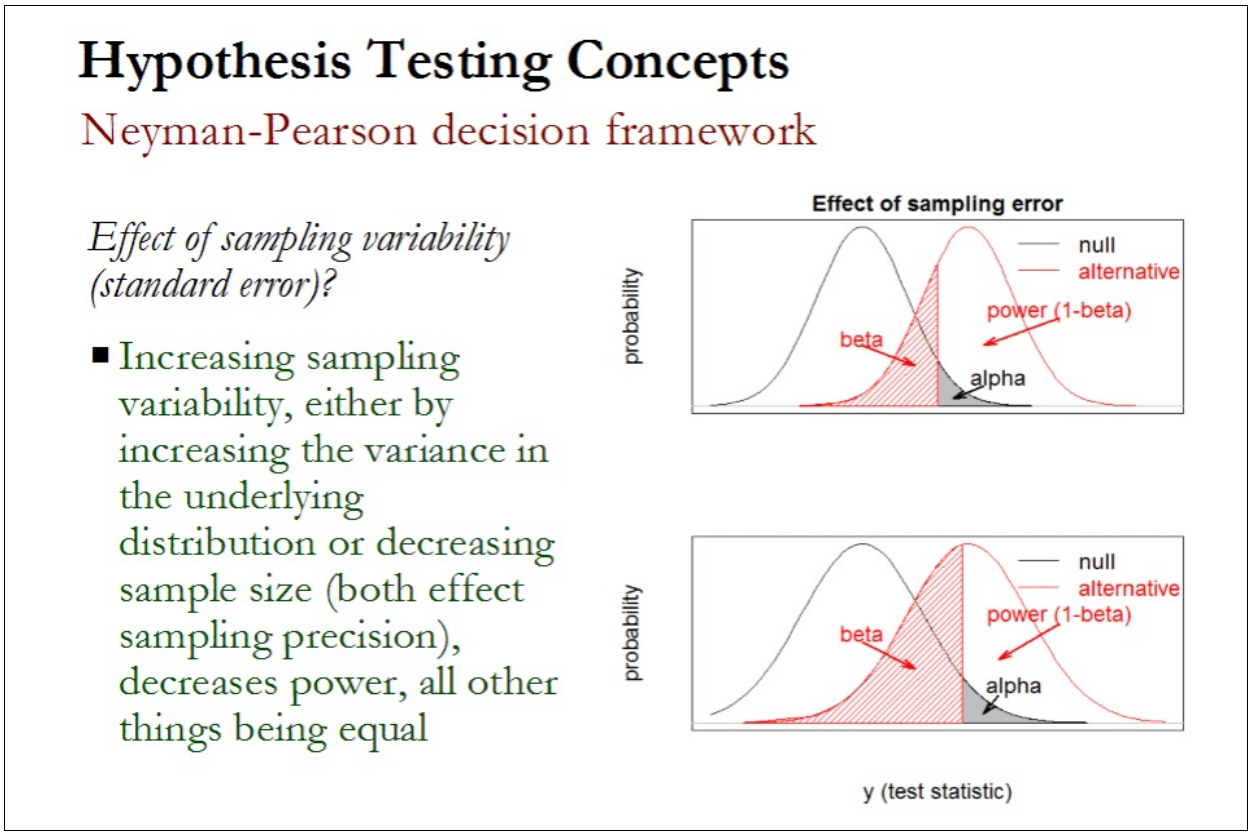
*Effect of alpha*?

- Increasing alpha, increases power, all other things being equal

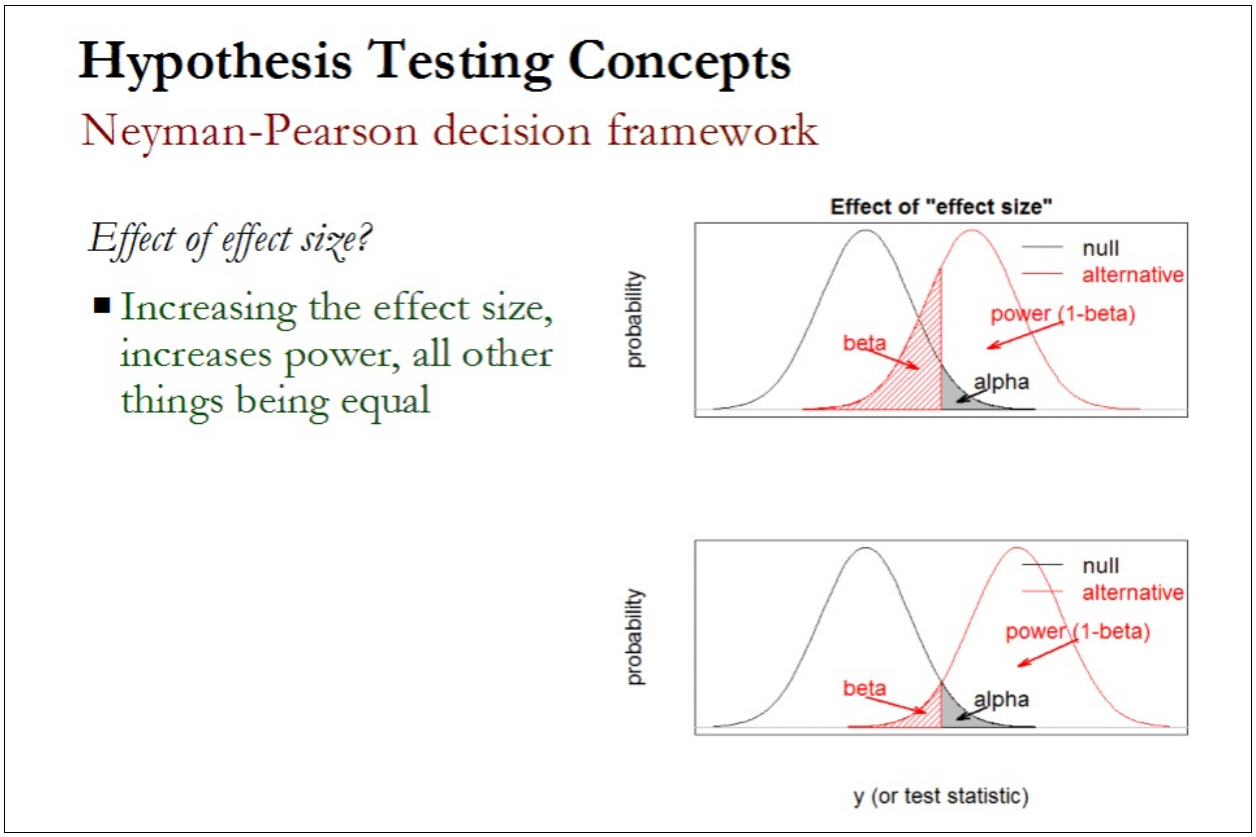



Once you understand the concept of power it is easy to see why power is affected by changes in *alpha*, sampling variability and effect size – all things that are partially or wholly under our control.

- Effect of *alpha* on power.–Based on the relationship between *alpha*, *beta* and power, it should be obvious why changing *alpha* affects power. If we increase alpha (i.e., choose a larger *p*-value to make it easier to reject the null hypothesis, but at the cost of increasing the Type I error rate), power will increase, all other things equal.




- Effect of sampling variability on power.–Similarly, it should be obvious why increasing sampling variability decreases power, all other things being equal. Note, sampling variability is influenced principally by two things: the inherent variability in the system, and the sample size. Our ability to precisely estimate any parameter is influenced by how variable the system is. If the system is highly variable, then for any fixed sample size, the precision of our estimate is going to be low – because any single sample is likely to vary wildly and not provide a very precise estimate of the true value. Unfortunately, often there is not much we can do to change the variability of the system. However, sampling precision is influenced by sample size, because for any given level of variability, the larger the sample size the better it represents the underlying population and thus the more confidence we have in our estimate of the population parameter. Fortunately, sample size is usually under our control, which is why the focus of most power analyses is on determining the sample size needed to achieve a desired level of power.




- Effect of "effect size" on power.–Effect size generally refers to the magnitude of difference between the null distribution and alternative distribution. Effect size can be measured in lots ways depending on the context of the power analysis, but in our example it is quite simply the difference between a slope of zero and whatever value we want to specify. For example, we might want to know the power associated with a slope of 0.006, an effect size of 0.006-0=0.006. What is the probability of rejecting the null hypothesis of zero slope if in fact the true slope is 0.006. As the effect size increases, power increases, all other things being equal. This makes sense, because as the difference between the null and alternative distribution increases, there is less and less overlap in the distributions, making it very unlikely that we would observe a value that would lead to a decision to accept the null when it was generated from the alternative distribution.