Hypothesis Testing

Inductive v. Deductive Reasoning

Inductive Inference: Small pieces of evidence are used to shape a larger theory.

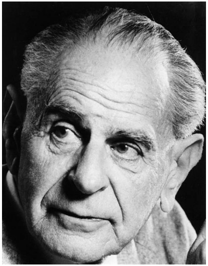
Deductive Inference:A larger theory is used to devise many small tests.

How Do we Derive Truth from Data?

1. Frequentist inference - correct conclusion drawn from repeated experiments
   - Null Hypothesis Tests
     - falsify a null hypothesis
   - Likelihood/Information Theoretic - evaluate weight of evidence
2. Bayesian - probability of belief that is constantly updated

Deductive vs. Inductive

Null Hypothesis Tests & Popper



Falsification of hypotheses is key!

A theory should be considered scientific if, and only if, it is falsifiable.

Deductive Reasoning and Null Hypothesis Tests

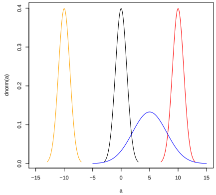
A null hypothesis is a default condition that we can attempt to falsify.

Common Uses of Null Hypothesis Tests

- Ho: Two groups are the same
- Ho: An estimated parameter is not different from 0
- Ho: The slopes of two lines are the same
- Etc...

Ho and Ha

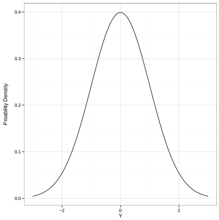
There are often many alternate hypotheses. Rejection of the null does not imply acceptance of any single alternative hypothesis.



Null Distributions

Null hypotheses are associated with null statistical distributions. For example, if Ho states that a value is normally distributed, but is not different from 0, the null distribtion is centered on 0 with some standard deviation.

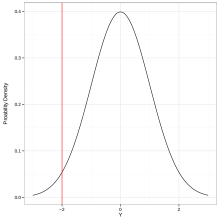
Null Distributions



Evaluation of a Test Statistic

We can use our data to calculate a test statistic that maps to a value of the null distribution. We can then calculate the probability of observing our data, or of observing data even more extreme, given that the null hypothesis is true.

Evaluation of a Test Statistic

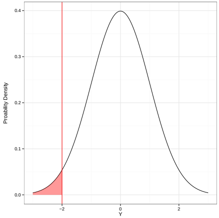


The P Value

P-value: The Probability of making an observation or more extreme observation given that the null hypothesis is true.

R. A. Fisher

The P Value



p=0.0227,Note - this is a one-tailed test!

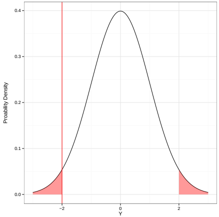
1-Tailed v. 2-Tailed Tests

1-Tailed Test: We are explicit about whether Ha implies that our sample is greater than or less than our null value.

2-Tailed Test: We are make no assumption about the sign or direction of our alternative hypotheses.

Two-Tailed P Value

```
## Error:  invalid argument to unary operator
```



p=0.0454 from `pnorm(-2)*2`

When should you use a 1-Tailed Test?

Exercise: Evaluate Support for Null Hypothesis

- Typically, the number of warts on a toad is Poisson distributed with a λ of 54
- You survey a lake suspected to contain high PAH levels. You pick up a toad, and it has 40 warts.
- What is your null hypothesis?
- What is the probability of making this observation, given your null?
- Challenge: How does your p value change with # of warts, say, from 1 to 108 warts?

Exercise: Evaluate Support for Null Hypothesis

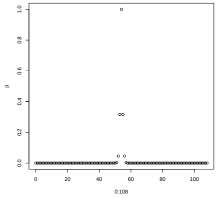
```
2 * ppois(40, 54)

## [1] 0.05755


# OR!
p <- 0
for (i in 1:40) {
    p <- p + dpois(i, 54)
}
p * 2

## [1] 0.05755
```

Exercise: Evaluate Support for Null Hypothesis

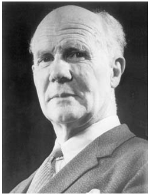


Exercise: Evaluate Support for Null Hypothesis

```
p <- 0
for (i in 0:54) {
    p[i + 1] <- 2 * pnorm(i, 54)
}
for (i in 55:108) {
    p[i + 1] <- 2 * pnorm(i, 54, lower.tail = F)
}
plot(0:108, p)
```

Neyman-Pearson Hypothesis Testing and Decision Making

Jerzy Neyman



Egon Pearson

Neyman-Pearson Hypothesis Testing

Rejection of a null hypothesis if the p-value is below some critical level - α

If p $\leq \alpha$ then we reject the null. There is *strong support* for the null to be falsified. This result is sometimes termed being statistically significant.

α is often 0.05, but, set it according to your a priori reasoning (including what you assume your power to be)

Statistical Significance is NOT Biological Sigficance.

Should we even use the word "significant"? Why or why not just talk about level of support for rejecting the null?

Types of Errors in a NHST framework

	Ho is True	**Ho is False**
Reject Ho	Type I Error	Correct OR Type S error
Fail to Reject Ho	–	Type II Error

- Possibility of Type I error regulated by choice of α
- Probability of Type II error regulated by choice of β
- Probability of Type III error is called δ

Type S Error

Correctly rejecting the null hypothesis for the wrong reason

This is a Type S, or Type III error - a mistake of sign. Often inherent in an experiment's design, or possible by change.
Can determine by mechanistic simulation or a redesigned study.

Power

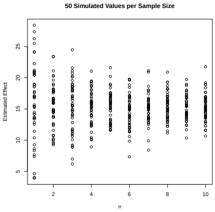
- If β is the probability of comitting a type II error, 1-β is the power of a test.
- The higher the power, the less of a chance of comitting a type II error.
- We typically want a power of 0.8 or higher.

Power via Simulation

We can assess the power of a test via simulation. We simulate a test statistic, and assuming a particular Ha is true, evaluate whether we falsely fail to reject Ho.

Sample Size and Power via Simulation

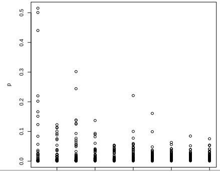
Ho is that the average effect of a drug on heart rate is 0. Actually, is speeds it up by 15 beats per minute. What is the effect of sample size of patients on power, assuming a SD of 6?



Sample Size and Power via Simulation

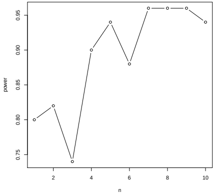
We can get the p value of each simulation using pnorm - and, remember, this is two-tailed!

```r
pvec <- pnorm(abs(vec), sd = simSD, lower.tail = FALSE) * 2
plot(pvec ~ n, ylab = "p")
```



Sample Size and Power via Simulation

Power is 1 - the fraction of those tests which p $\leq \alpha$. So, we loop over all sample sizes to get...



Sample Size and Power via Simulation

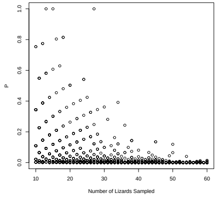
Power is 1 - the fraction of those tests which p $\leq \alpha$. So, we loop over all sample sizes to get...

```r
power <- rep(NA, 10)
for (i in 1:10) {
    nVec <- vec[which(n == i)]
    power[i] <- 1 - sum(nVec <= 0.05)/length(nVec)
}
plot(power ~ I(1:10), xlab = "n", ylab = "power", type = "b")
```

Exercise: Power and Simulation

- You allow lizards to choose to perch on a stick or remain on the ground
- Ho is that half will choose to perch. α is 0.05.
- Assuming that the probability that they will actually perch is 0.2, how is power affected by # of lizards?
- Challenge: How will this relationship be affected as you change alpha?

Exercise: Power and Simulation



Exercise: Power and Simulation