

Overview of Hypothesis Testing

Jon Brodziak

Northeast Fisheries Science Center

Classical View of Hypothesis Testing

- Begin with operational theory
- Assert a hypothesis (H) relevant to theory
- Identify observable prediction (P) from hypothesis
- Collect data for comparison with prediction
- Interpret evidence to disconfirm or confirm hypothesis

Logical Interpretation of Evidence

- Prediction is shown to be false
 - Evidence contrary to hypothesis is logically sufficient to disconfirm the hypothesis (modus tollens)
 - Reconsider theory and rethink hypothesis

Logical Interpretation of Evidence

- Prediction is shown to be true
 - Evidence supporting the hypothesis is not logically sufficient to confirm the hypothesis (affirming the consequent)
 - Eliminate all alternative hypotheses through experimental design and process of elimination (only H can explain P)
 - Hypothesis confirmation through elimination of alternatives is “strong inference”.
 - Results in which the effect E occurs in the presence of cause C but does not occur in the absence of C provide rigorous evidence for necessary and sufficient causation

Traditional Recipe for Hypothesis Testing

- Start with null hypothesis H_0 about some phenomenon or parameter (usually the opposite of what one wants to prove)
- Experimental data are collected to determine the parameter value
- A statistical test of the null hypothesis is conducted which generates a P-value
- The question of what the P-value means relative to H_0 is considered

What is a P-value (Not)?

- P is the probability that experimental results are due to chance. P close to 1 suggests results are due to chance and therefore it is safe to assume H_0 is true
- $1-P$ is the reliability of the experimental result or the probability of getting the same result if experiment is repeated
- P is the probability that H_0 is true

What is a P-value?

- P is the probability that experimental results are due to chance. P close to 1 suggests results are due to chance and therefore it is safe to assume H_0 is true
- $1-P$ is the reliability of the experimental result or the probability of getting the same result if experiment is repeated
- P is the probability that H_0 is true
- **WARNING:** Do not use these incorrect interpretations!
- Carver (1978) termed these interpretations to be “fantasies about statistical significance.”

What is a P-value?

- P is the probability of the observed data, or more extreme data, given that H_0 is true.
- $\Pr[\text{observed or more extreme data} \mid H_0 \text{ is true}]$
- Assumes that model is correct.
- Assumes that sampling is done randomly

Two Sample T-Test: Equal Variances

Test equality of means of populations X and Y

$$H_0: \mu_X = \mu_Y$$

Collect n samples from population X (X_1, \dots, X_n)
and m samples from population Y (Y_1, \dots, Y_m).

Two Sample T-Test: Equal Variances

Compute pooled sample variance

$$S_P^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

Two Sample T-Test: Equal Variances

Compute t-test statistic

$$t = \frac{\bar{X} - \bar{Y}}{S_P \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim T_{n+m-2}$$

Reject H_0 if $t > T_{n+m-2, \alpha/2}$ or $t < -T_{n+m-2, \alpha}$

Two Sample T-Test: Unequal Variances

Compute quotients of sample variances to sample sizes

$$q_X = \frac{S_X^2}{n} \text{ and } q_Y = \frac{S_Y^2}{m}$$

Compute the t-statistic

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{q_X + q_Y}}$$

Two Sample T-Test: Unequal Variances

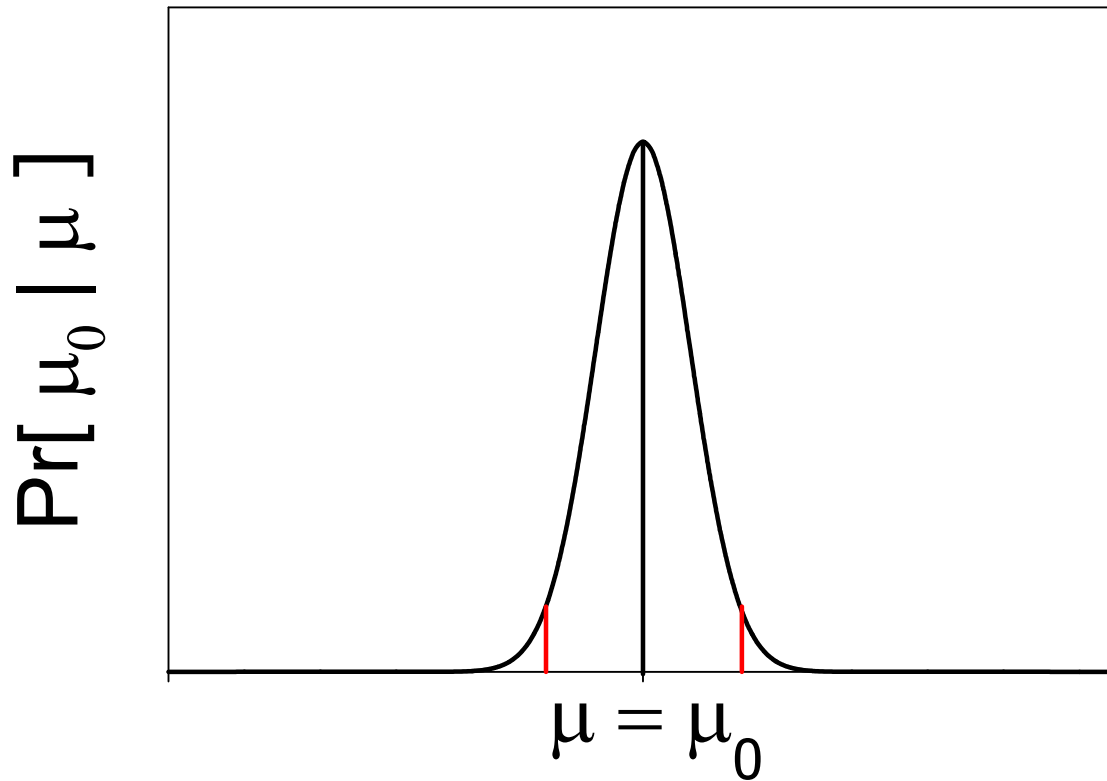
Compute the approximate degrees of freedom ν for the T distribution

$$\nu = \frac{\left(q_X + q_Y\right)^2}{\frac{q_X^2}{n-1} + \frac{q_Y^2}{m-1}}$$

Reject H_0 if $t > T_{\nu, \alpha/2}$ or $t < -T_{\nu, \alpha/2}$

What if the result is non-significant?

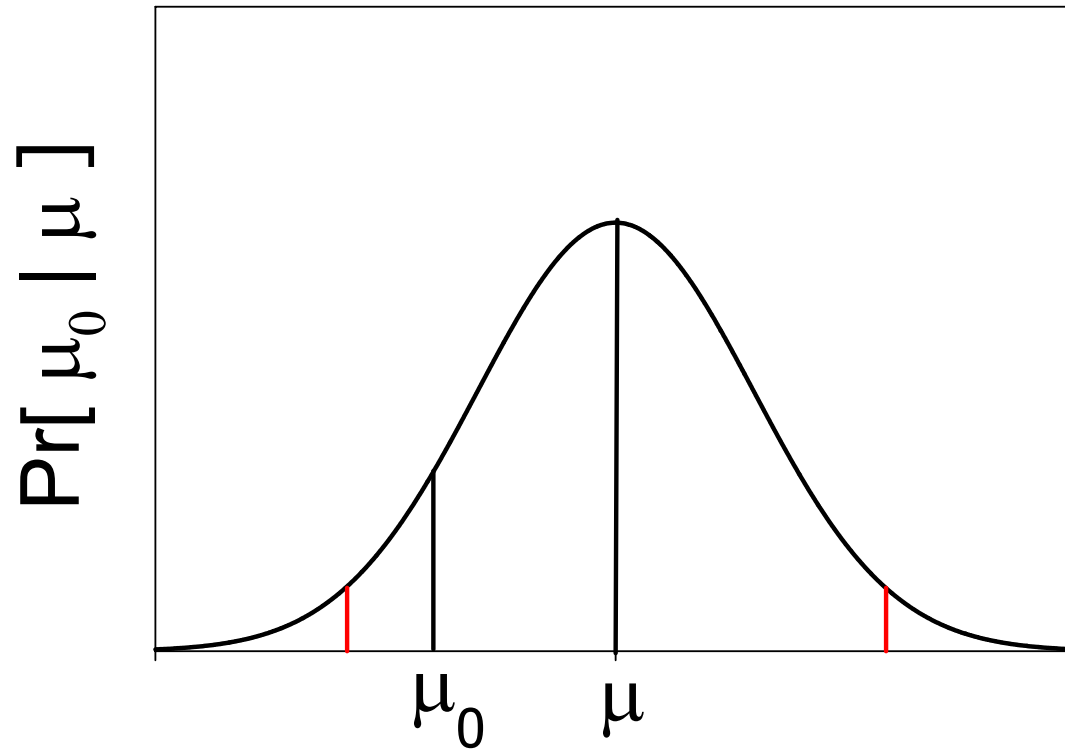
Failure to Reject $\mu = \mu_0$
Null Hypothesis is True



What if the result is non-significant?

Failure to Reject $\mu = \mu_0$

Lack of Power



P-Values Depend on Sample Size

Rewrite t-statistic as a function of mean difference and the sample size

$$t = \frac{\bar{D}}{S_D} \cdot \sqrt{n}$$

If $\bar{D} \neq 0$ then $t \uparrow$ as $n \uparrow$. And as $t \uparrow$, $P \downarrow$.

If null hypothesis is not absolutely true, one then can (in theory) make the P-value arbitrarily small by increasing sample size n.

Possible Effects of Sample Size on the Significance of Results

Practical Importance of Observed Difference	Not Significant	Significant
Not Important	N was okay	N was too big
Important	N was too small	N was okay

Focus on Estimating Effect Size and Its Precision

- Provide estimates of important parameters (e.g., effect size magnitude and sign) and measures of precision (preferably a confidence interval)
- Use standard errors for inference and to assess the precision of an estimator
- **Warning:** Do not provide a single or “naked” P-value to summarize your results—P-values are only moderately informative by themselves

Sample Size Needed to Achieve Desired Accuracy

- Suppose we specify an accuracy range d for the sample mean y
- We need this accuracy to apply with probability $1-\alpha$
- We need a sample size n large enough that the probability is at least $1-\alpha$ that $|y-\mu| < d$

Sample Size Needed to Achieve Desired Accuracy

If $y \sim N(\mu, \sigma^2)$, then d is given by

$$d = \frac{Z_{\alpha/2} \cdot \sigma}{\sqrt{n}}$$

If σ^2 is known, then take n to be

$$n \geq \left(\frac{Z_{\alpha/2} \cdot \sigma}{d} \right)^2$$

Sample Sizes for Testing Hypotheses

- Assume that we want to test the null hypothesis of no difference between group means
- TYPE I error rate is the chance of rejecting H_0 when H_0 is true, α
- TYPE II error rate is the chance of failing to reject a false null hypothesis, β
- Suppose we set $\alpha=0.01$. Then we reject H_0 only when the data are widely at variance with it.
- But for small α , the test may fail to reject H_0 even though H_0 is false (the alternative H_1 is actually true)
- Probability that a test will reject a FALSE null hypothesis given α is called its POWER, or $1-\beta$

State of Nature

Our Decision	H_0 is true	H_1 is true
Accept H_0	CORRECT DECISION	TYPE II Error β
Reject H_0	TYPE I Error α	CORRECT DECISION

Power of a Test Depends on

- Sample size, n
- Actual effect size, d
- Chosen significance level α

Consider alternative hypothesis carefully

- What is the direction of effect?
- What is the magnitude of effect?
- **WARNING:** Choosing a very small significance level α reduces the power $1 - \beta$ to detect real effects

Judging the Relative Risk of Error

- Suppose you choose $\alpha=0.001$ and this gives a power of $1-\beta=0.1$
- Then $\beta=0.9$ and the relative risk of rejecting a true H_0 is 900-fold more serious than accepting a false H_0 – this seems unbalanced
- Suppose instead you choose $\alpha=0.05$ and this gives a power of $1-\beta=0.8$
- Then $\beta=0.2$ and the relative risk of rejecting a true H_0 is 4-fold more serious than accepting a false H_0
- Use the ratio β/α to judge the relative risk of a TYPE I to a TYPE II error

Sample Size Needed to Achieve Desired Power

Compare means of two populations

$$H_1 : \mu_1 = \mu_2 + d$$

Estimate sample means y_1 and y_2
based on n_1 and n_2 samples and
compute T-statistic

$$t = \frac{y_1 - y_2}{N} \quad \text{where } N = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

Sample Size Needed to Achieve Desired Power

Note that

$$\begin{aligned}\beta &= \Pr(\textit{accept } H_0 \mid H_1 \textit{ true}) = 1 - \textit{power} \\ &= \Pr(t < Z_\alpha \mid H_1 \textit{ true})\end{aligned}$$

Therefore the power of the test satisfies

$$\begin{aligned}1 - \beta &= \Pr(\textit{reject } H_0 \mid H_1 \textit{ true}) \\ &= \Pr(T > Z_\alpha \mid H_1 \textit{ true})\end{aligned}$$

Sample Size Needed to Achieve Desired Power

Note that

$$1 - \beta = \Pr\left(T - \frac{d}{N} > Z_{\alpha} - \frac{d}{N} \mid H_1 \text{ true}\right)$$

Or alternatively

$$Z_{\alpha} - \frac{d}{N} = Z_{1-\beta} = -Z_{\beta}$$

Sample Size Needed to Achieve Desired Power

Solve for n_1+n_2

assuming wlog that $n_1 = g \cdot n_2$ for some constant $g > 0$

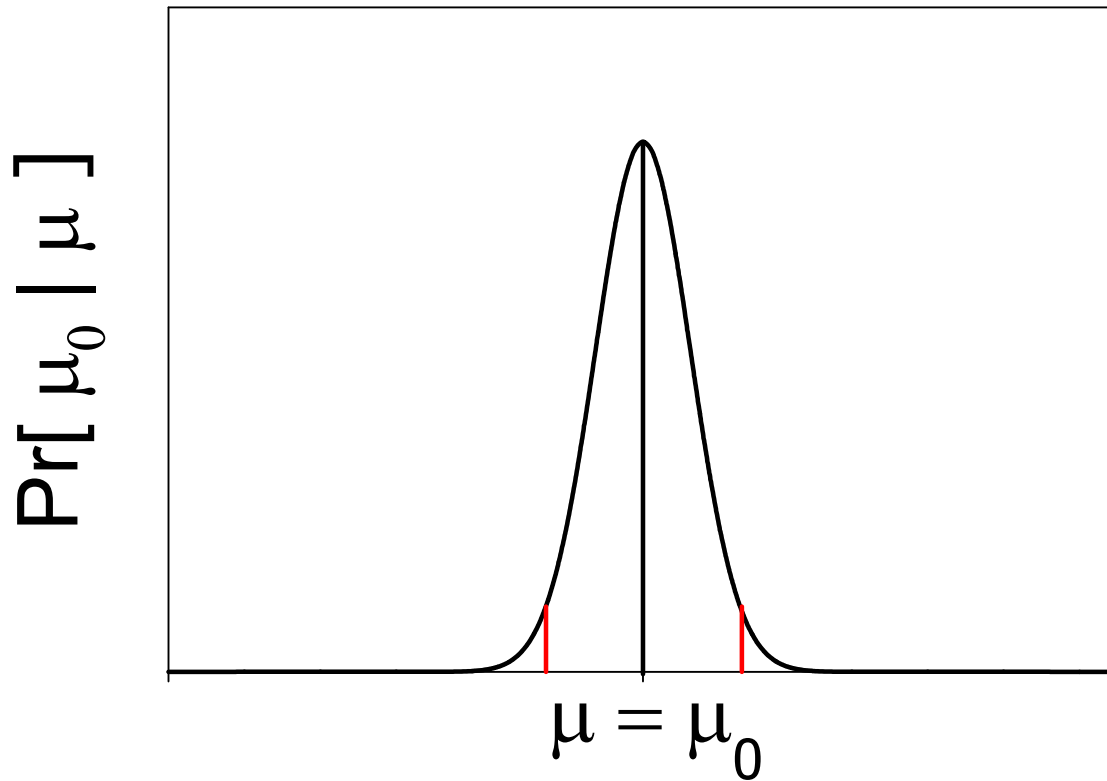
$$n_1 + n_2 = \left(Z_\alpha + Z_\beta \right)^2 \sigma^2 \frac{(g + 1)^2}{g \cdot d^2}$$

Warning: Avoid Post Hoc Power Calculations

- Suppose we have a non-rejected null hypothesis
- Some advocate computing the power of the test for the observed value of the test statistic
- Since the observed significance level of a test also determines the observed power, reporting observed power adds nothing to the interpretation of results
- Citing observed power AFTER observing the P-value provides no additional information to interpret the test results – it is devoid of meaning
- See Hoenig, J. and D. Heisey. 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. *American Statistician*. 55:19-24

What if the result is non-significant?

Failure to Reject $\mu = \mu_0$
Null Hypothesis is True



What if the result is non-significant?

Failure to Reject $\mu = \mu_0$

Lack of Power

