STAT 518 --- Section 2.3: Hypothesis Testing

• Often in scientific studies, the researcher presents a specific claim about the population.

• We gather data, and based on these data determine whether or not the claim appears to be true.

Example 1: We gather experimental data to determine whether drug A is equally effective, on average, as drug B.
Example 2: We gather survey data to test the claim that no fewer than 50% of registered voters support the governor's latest policy.
Example 3: We gather observational data to determine whether a verbal test score distribution for females matches the corresponding distribution for males.

• Statistical hypotheses are stated in terms about the population (possibly, about one or more parameters).

• The research hypothesis (or alternative hypothesis, denoted by H_1 or H_a) represents a theory that the researcher suspects, or seeks evidence to "prove."

• The null hypothesis (denoted by H_0) is the negation (opposite) of H_1.

• H_0 often represents some "previously held belief," "status quo," or "lack of effect."

• If we gather a set of sample data and it would be **highly unlikely** to observe such data if H_0 were true, then we have evidence against H_0 and in favor of H_1.

• We must select a **test statistic**: a function of the data whose value indicates whether or not the data agree with H_0.

• We formulate a **decision rule**, which tells us which values of the test statistic lead us to **reject** H_0.

• Based on the data from our random sample, we calculate the test statistic value and use the decision rule to decide whether or not to reject H_0.

Example 2 Hypotheses:

$H_0: p \geq 0.5$

$H_1: p < 0.5$

• Suppose we will select a random sample of 20 voters and ask each whether he/she agrees with the policy:

Test statistic: T = the number in the sample who agree with the policy

Decision rule: Reject H_0 if the test statistic is sufficiently small.

Note if H_0 is true, T has binomial distribution with $n=20$, $p=0.5$

Let's say 5 of the 20 agree with the policy. If *p* were 0.5, then

$P(T \leq 5) =$ 0.0207 (Table A3)

- **Is this unlikely enough to cause us to reject the notion that *p* is at least 0.5?**

Types of Hypotheses

- **A hypothesis is <u>simple</u> if it implies only one possible probability function for the data.**

- **A hypothesis is <u>composite</u> if it implies numerous possible probability functions for the data.**

<u>Example 2 above</u>: Simple or composite hypotheses?

Composite for both H_0 and H_1

- **A** simple **hypothesis in the case of Example 2 would be:**

$H_0: p = 0.5$

Critical Region

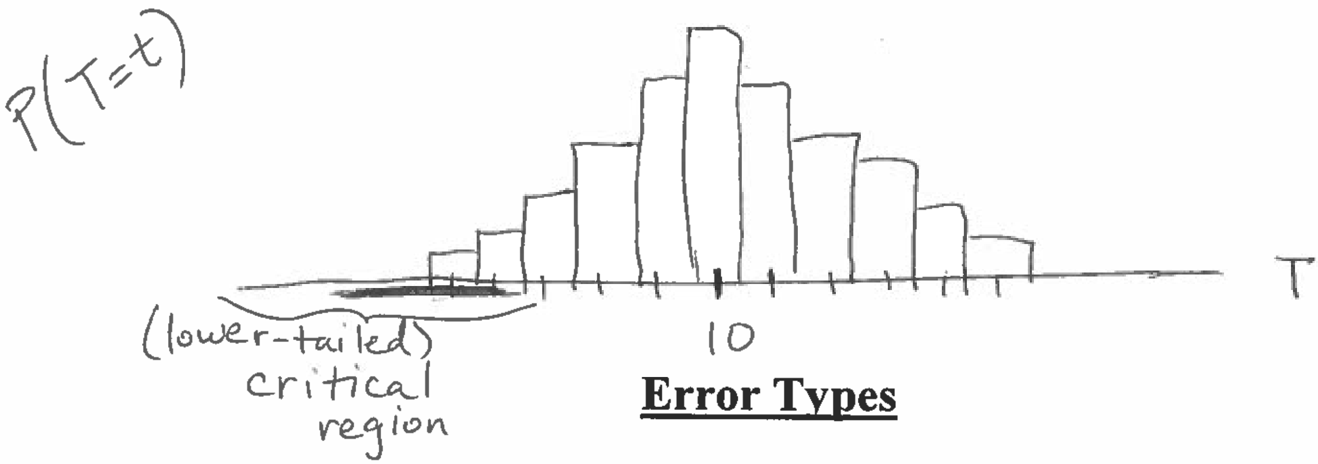
- **The <u>critical region</u> (or** rejection **region) is the set of all test statistic values that lead to rejection of the null hypothesis.**

- **Our <u>decision rule</u> establishes the critical region.**

• **If the critical region contains only small values OR only large values of the test statistic, we have a** one-tailed **test.**

• **If the critical region contains BOTH small and large values of the test statistic, we have a** two-tailed **test.**

Example 2 above:



Error Types

• **There are two types of incorrect decisions when performing a hypothesis test.**

• **We could make a Type I error: Rejecting H_0 when it is in fact true.**

• **We could make a Type II error: Failing to reject H_0 when it is in fact false.**

• **The level of significance (denoted α) of the test is the maximum allowable probability of making a Type I error.**

• We typically let α be some small value and then determine our corresponding critical region based on the null distribution of the test statistic.

• The <u>null distribution</u> ~~is the distribution~~ of the test statistic is its probability distribution when the null hypothesis is assumed to be true.

Back to Example 2. What is α if our decision rule is "Reject H_0 if $T \leq 6$"? What is the P[Type I error]?

Null distribution of T: Binomial $(n=20,\ p=0.5)$

$$\alpha = P[\text{Reject } H_0 \mid H_0 \text{ true}]$$

$$= P[T \leq 6 \mid T \sim \text{Bin}(20, 0.5)]$$

$$\Rightarrow \alpha = 0.0577 \text{ from Table A3.}$$

<u>Power</u>

$\beta = P[\text{Type II error}]$

• The <u>power</u> (denoted $1 - \beta$) of a test is the probability of rejecting H_0 when H_0 is <u>false</u>.

• If H_1 is simple, the power is a single number.

• If H_1 is composite, the power depends on "how far away" the truth is from H_0 (more later).

P-value

- **Given observed data and the corresponding test statistic t_{obs}, the p-value is the probability of seeing a test statistic as or more favorable to H_1 as the t_{obs} that we did see.**

Examples

Lower-tailed test: P-value = $P(T \le t_{obs})$

using the null distribution of T.

Upper-tailed test: P-value = $P(T \ge t_{obs})$

using the null distribution of T.

Two-sided test: P-value defined to be:

$$2 \times \left[\min \left\{ P(T \le t_{obs}), P(T \ge t_{obs}) \right\}\right]$$

Example 2 again: P-value was $P(T \le t_{obs})$

$= P(T \le 5)$ where $T \sim \text{Bin}(20, 0.5)$

$= 0.0207$ from Table A3.

If $\alpha = 0.0577$, then we reject $H_0: p \ge 0.5$

since our P-value $\le \alpha$.

Note: We reject H_0 whenever P-value $\le \alpha$.

Section 2.4: Properties of Hypothesis Tests

• **Often there are multiple test procedures we could use to test our hypotheses of interest.**

• **How to decide which is the best to use?**

• **Note that some tests require certain assumptions about the data.**

Example: Classical t-test about μ:

Requires data follow a normal distribution.

• **A test that makes less restrictive assumptions may be preferred to one whose assumptions are more stringent.**

• **If the assumptions of a test are not in fact met by the data, using the test may produce invalid results.**

Properties of Tests

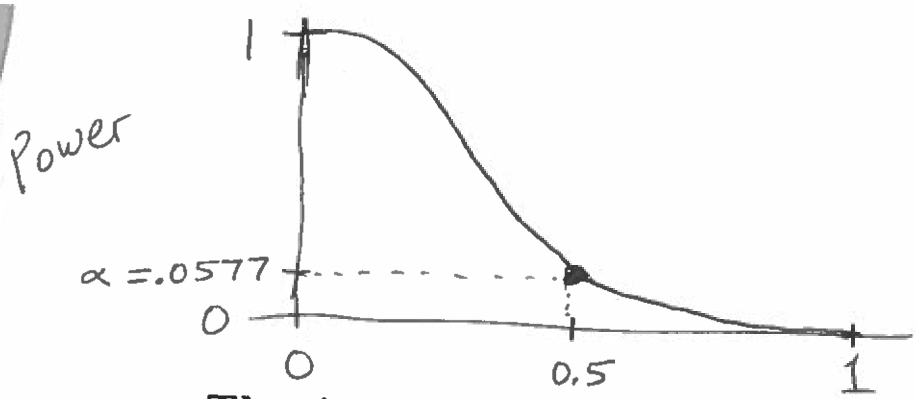
Power Function: Often the hypotheses H_0 and H_1 are written in terms of a parameter of interest.

• **The power function of a test describes P[Reject H_0] as a function of the parameter value.**

Example 2 again: Note *p* could be between 0 **and** 1**.**

H_0: $p \geq 0.5$ **H_1:** $p < 0.5$

Using decision rule:
Reject H_0 when $T \le 6$



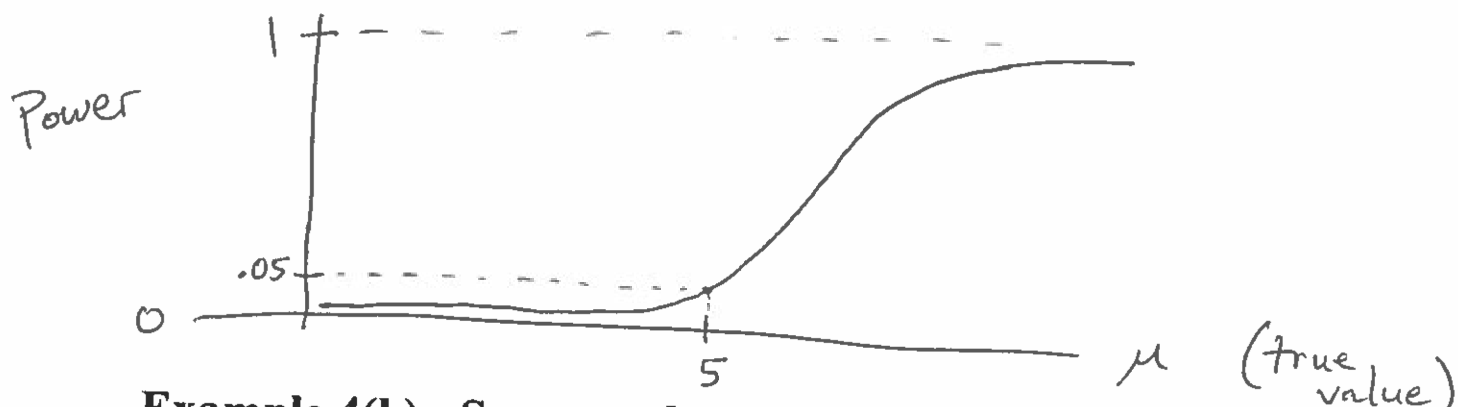
• The **significance level** is the **maximum** value of the power function **over the region** corresponding to H_0.

Example 4(a): Suppose we test H_0: $\mu \le 5$ vs. H_1: $\mu > 5$ based on 100 observations from a $N(\mu, 1)$ population, using $\alpha = 0.05$.

• **We use a** z-test **: Reject H_0 if**

$$z = \frac{\bar{X} - 5}{1/\sqrt{100}} > 1.645$$

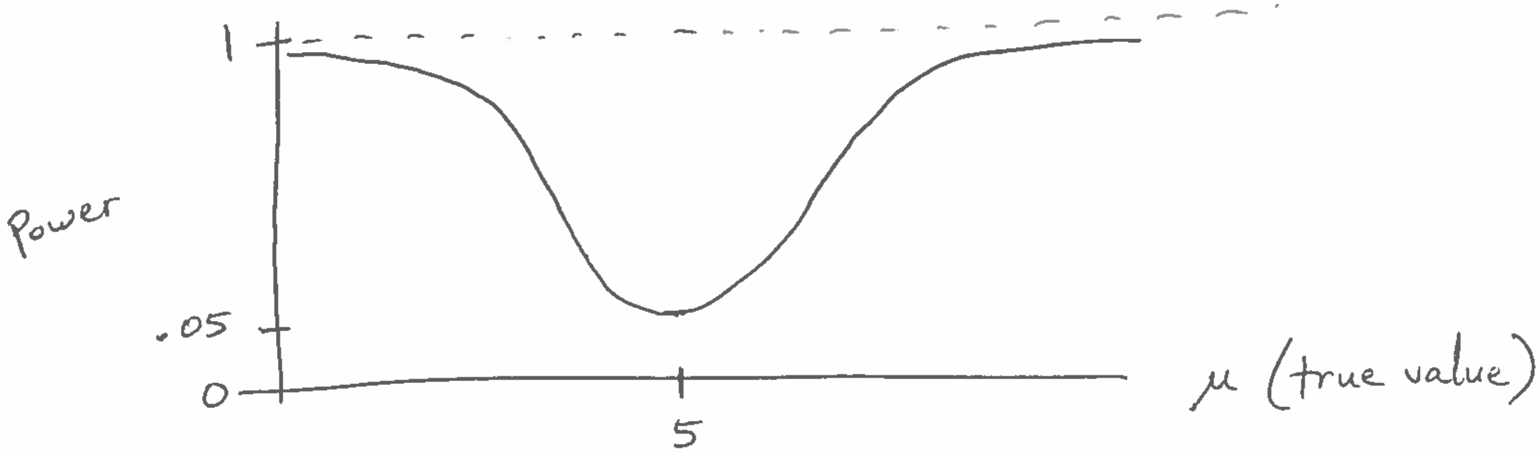
Power function:



Example 4(b): Same as above, but we test H_0: $\mu = 5$ vs. H_1: $\mu \neq 5$.

• **Our test is: Reject H_0 if** $|z| > 1.96$

Power function:



• A test is __unbiased__ if P[Reject H_0] is always at least as large when H_0 is false as when H_0 is true.

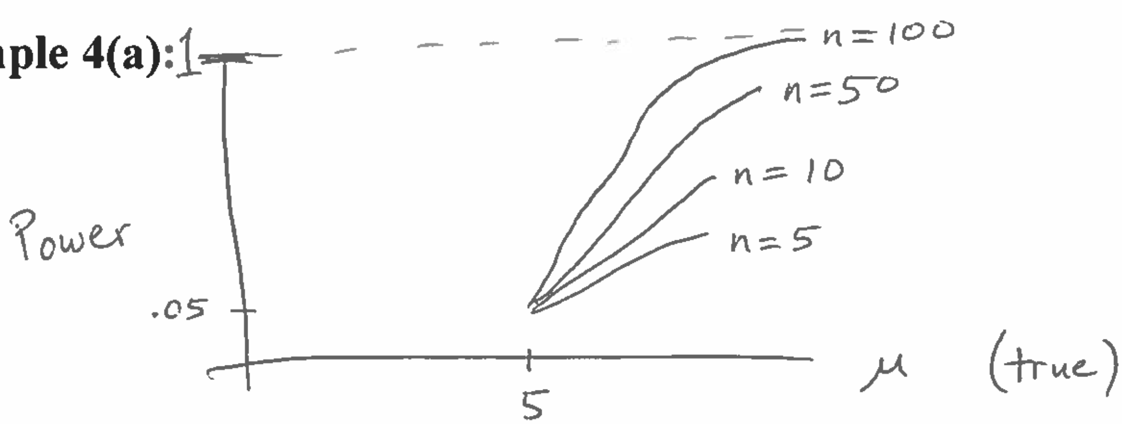
__Example 2__: Unbiased
__Example 4(a)__: Unbiased
__Example 4(b)__: Unbiased

• We would like our test to have more __power__ to reject a false H_0 when our sample size grows larger.

• A test (actually, sequence of tests) is __consistent__ if for __every__ parameter value in H_1, the power $\longrightarrow 1$ as $n \to \infty$

• This assumes the level of significance of the tests in the sequences does not exceed some fixed α.

Example 4(a):



Calculating Power if both H_0 and H_1 are Simple

- **Recall Example 2, but now suppose the hypotheses are**

$$H_0: p = 0.5 \quad \text{vs.} \quad H_1: p = 0.3$$

and suppose again that our decision rule is "Reject H_0 if $T \le 6$" where T = number of voters out of the 20 sampled who agree with the governor's policy.

- **We have already calculated that our significance level of this test is**

$$\alpha = P[\text{Reject } H_0 \mid H_0 \text{ true}]$$
$$= P[T \le 6 \mid T \sim Bin(20, 0.5)]$$
$$= 0.0577 \quad (\text{Table A3})$$

- **When both H_0 and H_1 are simple hypotheses, the power will be a single number, which we can easily calculate:**

$$\text{Power} = P[\text{Reject } H_0 \mid H_0 \text{ false}]$$
$$= P[T \le 6 \mid T \sim Bin(20, 0.3)]$$
$$= 0.6080 \quad (\text{Table A3})$$

- **If we change our decision rule to "Reject H_0 if $T \le 5$", what happens to the significance level?** ↖ more stringent

$$\alpha = P[T \le 5 \mid T \sim Bin(20, 0.5)] = 0.0207$$

What happens to the power?

$$\text{Power} = P[T \le 5 \mid T \sim Bin(20, 0.3)] = 0.4164$$

<u>Tradeoff</u>: P[Type I error] goes down, but power also goes down.

Comparing Two Testing Procedures

- **Suppose we have two procedures T_1 and T_2 to test H_0 and H_1.**

- **Assume the significance level α and the power are the same for each test.**

- **The test requiring the smaller sample size to achieve that power is more efficient.**

- **The** relative efficiency **of T_1 to T_2 is** $\frac{n_2}{n_1}$

where n_1 = the required sample size for T_1
n_2 = the required sample size for T_2.

- **If eff(T_1, T_2) > 1, then and T_1 is** more **efficient than T_2.**

- **If H_1 is composite, the relative efficiency may be different for each parameter value in the alternative (in H_1) region.**

- **A measure of efficiency that does not depend on α, power, or the alternative is the asymptotic relative efficiency (A.R.E.) (or Pitman efficiency).**

- **If we can find a relative efficiency n_2/n_1 such that this ratio approaches a constant as $n_1 \to \infty$ (no matter which fixed α and power are chosen), then the limit of n_2/n_1 is the A.R.E. of T_1 to T_2.**

• **We often use the A.R.E. to measure which test is superior.**

• **Although A.R.E. compares tests based on an infinite sample size, it works fairly well as an approximation of relative efficiency for practical sample sizes.**

• **The** actual **significance level of a test is the probability that H_0 is actually rejected (if H_0 is true).**

<u>Conservative Test</u>: A test is <u>conservative</u> if the actual **significance level is** smaller **than the stated (or <u>nominal</u>) significance level.**

nominal level

<u>Example 2 again</u>: Suppose our stated $\alpha = 0.05$.

Decision rule should be:

Reject H_0 if $T \leq 5$ (more stringent)

Actual significance level is: $0.0207 <$ stated α

$\Rightarrow$ Test is <u>conservative</u>.

(drawback : <u>less</u> <u>power</u>)

Section 2.5: Nonparametric Statistics

- **Parametric methods of inference depend on knowledge of the underlying population distribution.**

Example 4: We assumed the data followed a normal **distribution.**

- **We cannot be certain of the distribution of our sample of data.**

- **We can use preliminary checks (plots, tests for normality) to determine whether the data might reasonably be assumed to come from a normal distribution.**

- **The classic tests learned in STAT 515 are efficient and powerful when the data are truly normal.**

Robust Methods

- **A robust method is one that works fairly well even if one of its assumptions is not met.**

- **The t-tests (one- and two-sample) are robust to the assumption of normality.**

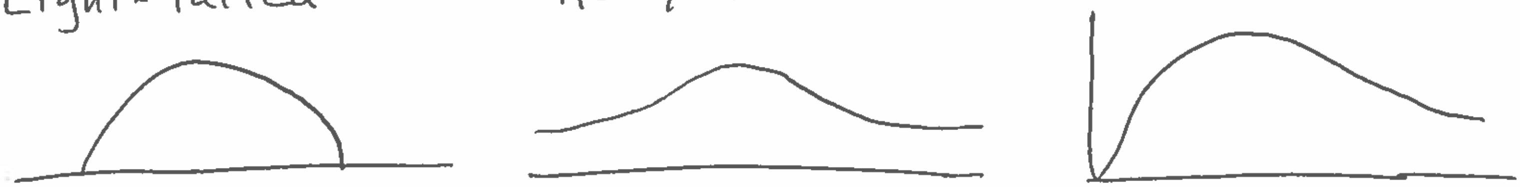
- **Even if the data are somewhat non-normal, the actual significance level will be close to the nominal significance level.**
- **However, is the t-test powerful in that case?**

• Parametric procedures tend to:

- have good power when the population is light-tailed
- have low power when the population is heavy-tailed
- have low power when the population is skewed

Pictures:

Light-tailed   Heavy-tailed   Skewed



• A sample with outliers is a sign of a possibly heavy-tailed **population distribution.**

• Many classic parametric procedures are asymptotically distribution-free:

- As the sample size gets larger, the method gets more robust.

- When the sample size is extremely large, the type of population distribution may not matter at all.

• The t-tests are asymptotically distribution-free

because of the central limit theorem**.**

• Still, for small to moderate sample sizes, being asymptotically distribution-free is irrelevant: We should pick the procedure that is most powerful and efficient.

Nonparametric Methods

• **Definition**: A statistical method is called **nonparametric** if it meets at least one of these criteria:

(1) The method may be used on data with a nominal measurement scale.
(2) The method may be used on data with an ordinal measurement scale.
(3) The method may be used on data with an interval or ratio measurement scale, where the form of the population distribution is unspecified. (distribution-free)

Example 2 data: Each observation is Yes or No
$\Rightarrow$ nominal data

Criterion (1) is satisfied by our binomial-type test.

Example 3 data: If we do not claim to know the population distributions of the test scores:

A nonparametric test satisfying criterion (3) may be used.