Hypothesis Testing

Recall: So far, we have discussed point estimates - finding a "best guess" for an unknown population parameter and confidence intervals - plausible values for the unknown population parameter. The objective of a statistical hypothesis test is to decide which one of two contradictory claims about the population parameter is correct.

DEFINITION: A statistical hypothesis is a claim about a population parameter θ (such as the mean μ or a proportion p). The NULL HYPOTHESIS, denoted by H_0 is the claim that is initially assumed to be true. The ALTERNATIVE HYPOTHESIS, denoted by H_a is an assertion that is contradictory to H_0.

If the observed data is plausible (has high probability) under the null hypothesis assumption, we will accept this claim as true. If the observed data has very low probability under the null hypothesis and higher probability under the alternative hypothesis, we will REJECT the null hypothesis in favor of the alternative.

Testing procedure: A statistical hypothesis test consists of several components.

1. The null hypothesis statement and the alternative statement. The null hypothesis should always be phrased as an equality (e.g., $H_0 : \mu = 0$, or $H_0 : p = 0.5$, or $H_0 : \theta = 0.2$). The alternative can be phrased as an equality (e.g., $H_a : \mu = 3$) or an inequality (e.g., $H_a : \mu \neq 0$ or $H_a : p > 0.5$, or $H_a : \theta < 0.2$).

2. A TEST STATISTIC. This is a function whose value can be computed from the sample data and whose (theoretical) distribution is known if the null hypothesis H_0 is true. The decision whether to accept or reject H_0 is based on the value of the test statistic computed from the data.

3. A REJECTION REGION - the set of all test statistic values for which the null hypothesis H_0 will be rejected.

Errors: There are two possible errors that can be made in hypothesis testing:

- Rejecting the null hypothesis H_0 when it is true (type I).
- Failing to reject the null hypothesis H_0 when it is false (type II).

Ideally, one would want to keep the probabilities of both these errors as small as possible. However, the error probabilities are related and if one error probability is made smaller the other one usually will increase. The choice of rejection region determines the probabilities of both a type I and type II error.

same α in the confidence interval

DEFINITION: The probability of a type I error α is called the (significance) LEVEL of the test. The probability of a type II error is usually denoted by β. The quantity $1 - \beta$ represents the test's ability to correctly reject a false null hypothesis and is called the POWER of the test.

Example: Two bags contain 2 white and 2 black (bag 1) and 1 white and 3 black marbles (bag 2), respectively. A bag is chosen at random and from that bag two marbles are selected at random.

(a) Let p denote the proportion of white marbles in the selected bag. Formulate the null hypothesis and alternative hypothesis for this example.

H_0=

(b) Based on the colors of the two marbles drawn, formulate a test statistic function. What is the distribution of this test statistic function if H_0 is true?

(c) What are the values of the test statistic function that make least likely that the selected bag is bag 1? Define the rejection region for this hypothesis test.

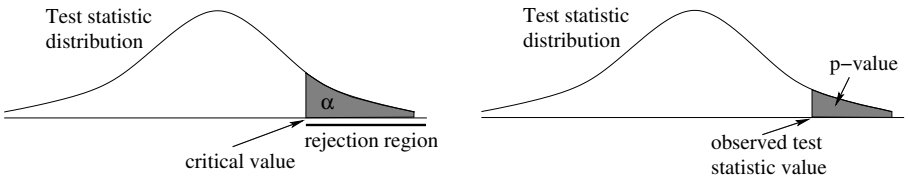
(d) For the above rejection region, compute the probability that the null hypothesis is rejected when it is, in fact, true ($\alpha = P(\text{type I error})$).

(e) For the above rejection region, compute the probability that the null hypothesis is not rejected when it is, in fact, false ($\beta = P(\text{type II error})$).

Poll Question 12.1- 12.3

P-Values

One way to report the results of a hypothesis test is to say whether or not the test statistic value fell into the rejection region and subsequently, whether or not the null hypothesis was rejected at a specified level of significance α. This yes/no decision does not convey any information about how soundly the null hypothesis was rejected. *Where* in the rejection region did the observed test statistic value fall?



DEFINITION: Suppose the null hypothesis H_0 is, in fact true. The p-value is the probability to observe a test statistic value at least as contradictory to H_0 as the computed value by random chance due to the selection of the sample. To compute this probability, we use the distribution of the test statistic function that is valid if the null hypothesis is true.

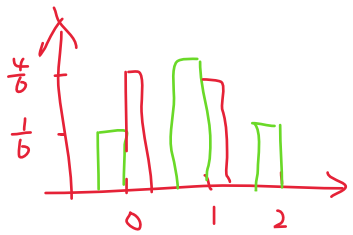
Poll Question 12.4

Example: (cont.)

Suppose the two selected marbles in the previous example are black and white. Compute the p-value for the null hypothesis that the proportion of white marbles in the bag is $p = 0.5$.

H0: p=0.5
Ha: p=0.25

X=# of white
Observed X=1



p value=P(extreme if H0 is true)
=P(X<=1 if H0 is true)

Poll Question 12.5

DEFINITION: If a p-value is smaller than the significance level α, then the corresponding value of the test statistic falls into the rejection region and the null hypothesis will be rejected.

$$p \leq \alpha \quad \Rightarrow \quad \text{reject } H_0$$

If the p-value is large, then it is quite likely to see data such as the observed by random chance if the null hypothesis were true and H_0 will not be rejected.

$$p > \alpha \quad \Rightarrow \quad \text{fail to reject } H_0$$

IN GENERAL: Follow this procedure whenever you conduct a statistical hypothesis test:

0. Pick a reasonable value of α (unless somebody else has already picked it for you).
1. Identify the parameter of interest in the problem (e.g., mean μ or proportion p, etc.)
2. Formulate the null hypothesis H_0 and the alternative hypothesis H_a.
3. Select a test statistic and compute the test statistic value for the sample data.
4. (a) EITHER: Find the rejection region for your type of alternative and level of α. Determine whether or not your test statistic value falls into the rejection region.

   (b) OR: (better!) compute the p-value for your observed test statistic value. Compare the p-value to α. Reject H_0 if $p \leq \alpha$, do not reject H_0 if $p > \alpha$.
5. Draw a conclusion and decide whether or not to reject H_0. Your conclusion should always be formulated as a sentence (not a formula) and be worded in the context of the original example.

p：probability
α：
x：value
c：critical value
x, c have same unit
p, α have same unit

Test for a Population Proportion

Let p denote the proportion of individuals in a population who possess a certain characteristic (successes). A random sample of size n is selected from the population and the proportion $\hat{p}$ of successes in the sample is observed. We want to use this quantity to decide whether to accept or reject a statement about the population parameter p.

$\hat{p} = \frac{x}{n} = \frac{\#\text{ success}}{n} = \frac{1}{n}\sum x_i = \bar{x}$

LARGE SAMPLE TESTS

If the sample size n is large ($np_0 \geq 10$ and $n(1-p_0) \geq 10$), then the test statistic for the hypothesis test

$$H_0 : p = p_0$$

has approximately a Normal distribution

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim \text{Normal}(0, 1).$$

Null hypothesis: $H_0 : p = p_0$

Test statistic: $z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$

Alternative Hypothesis	Rejection region at level α
$H_a : p > p_0$	$z \geq z_\alpha$ (upper-tailed test)
$H_a : p < p_0$	$z \leq -z_\alpha$ (lower-tailed test)
$H_a : p \neq p_0$	$z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$ (two-tailed test)

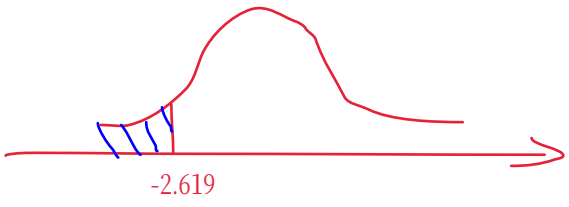
Example: Prevnar is a vaccine for meningitis usually given to infants. In a clinical trial, Prevnar was given to 710 children, of whom 72 experienced a loss of appetite. Competing medications cause about 13.5 percent of children to experience a loss of appetite. Can we conclude that the percentage of children experiencing a loss of appetite from Prevnar is significantly less than for other medications?

choose a $\alpha = 0.05$ (you can choose a other proper α)

1. p= proporthion of kids who would have loss of appetite from Prevnar
2. H0: p= 0.135 Ha: p<0.135
3.

p0=0.135
n=710
$\hat{p}$ =72/710

Z=-2.619



p=P(Z<-2.619)=0.00441< α

Reject H0 Conclusion:
The proportion of children experiencing loss of appetite from Prevnar is significantly lower than for other medications

Tests for a Population Mean

CASE I: Normal Population with known σ

Let $x_1, \ldots, x_n$ be observations from a Normal population with unknown mean μ and known variance σ^2. According to the Central Limit Theorem, the sample mean has a Normal distribution with mean μ and variance σ^2/n.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{ Normal}(0, 1)$$

Suppose we want to test the null hypothesis that $H_0 : \mu = \mu_0$ (some specified value).

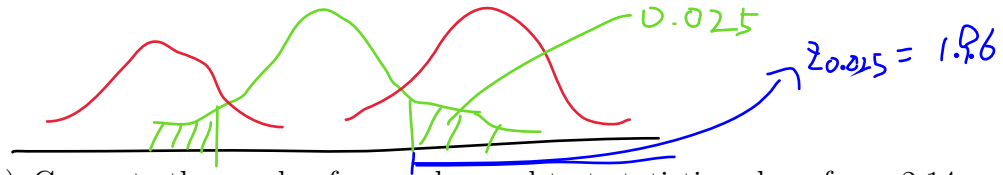
In this case the function

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

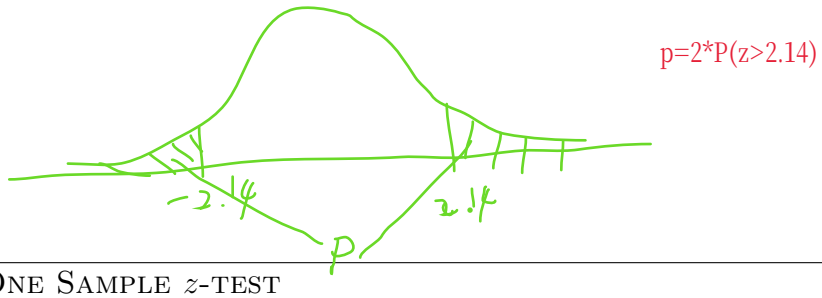
will become our test statistic function. The distribution of the test statistic is known if H_0 is true. This allows us to find rejection regions for specified levels α or to compute p-values for certain values z of the test statistic function.

Example: Suppose the null hypothesis is $H_0 : \mu = 5$ with alternative $H_a : \mu \neq 5$. both side

(a) Find the rejection region for the test statistic for level $\alpha = 0.05$.



(b) Compute the p-value for an observed test statistic value of $z = 2.14$.



ONE SAMPLE z-TEST

Null hypothesis $H_0 : \mu = \mu_0$

Test statistic: $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

Alternative Hypothesis	Rejection region at level α
$H_a : \mu > \mu_0$	$z \geq z_\alpha$ (upper-tailed test)
$H_a : \mu < \mu_0$	$z \leq -z_\alpha$ (lower-tailed test)
$H_a : \mu \neq \mu_0$	$z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$ (two-tailed test)

Poll Question 12.6

Example: Light bulbs of a certain type are advertised as having an average lifetime of 750 hours (with standard deviation $\sigma = 5.4$ assumed known). The price of these bulbs is very favorable, so a potential customer has decided to go ahead with a purchase unless is can be conclusively demonstrated that the true average lifetime is smaller than advertised. A random sample of 50 bulbs was selected, the lifetime of each bulb determined and the sample statistic $\bar{x} = 738.44$ computed.

mean case, σ known
choose α =0.05
1. μ =mean average lifetime of light bulb
H_0= μ =750, H_a: μ < 750
2.n=50, σ =50
Calculate Z=(738.44-750)/(5.4/sqrt(50))=-15.137
3.p=P(Z<-15.137)=0
4.Reject H_0
5.Conclusion:the average lifetime is significantly shorter than 750

CASE II: Large-Sample Tests

If a sample is large, then the population standard deviation σ may be replaced by its point estimate s without changing the distribution of the sample mean:

$$\text{Test statistic:} \qquad Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim \text{Normal}(0, 1)$$

CASE III: Normal Population Tests

If the sample size n is not large but the distribution of X is Normal, then replacing the population variance σ with the sample variance s changes the distribution of the sample mean from Normal to the t-distribution.

$$\text{Test statistic:} \qquad T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(df = n - 1)$$

The testing procedure in this case is very similar to the z-test procedure, except that the rejection region (or p-value computation) now depends on the t-distribution instead of the Normal distribution.

ONE SAMPLE t-TEST

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$

Alternative Hypothesis	Rejection region at level α
$H_a : \mu > \mu_0$	$t \geq t_{\alpha,n-1}$ (upper-tailed test)
$H_a : \mu < \mu_0$	$t \leq -t_{\alpha,n-1}$ (lower-tailed test)
$H_a : \mu \neq \mu_0$	$t \leq -t_{\alpha/2,n-1}$ or $t \geq t_{\alpha/2,n-1}$ (two-tailed test)

Example: Past data have shown that if parking meters in a very small town are emptied every 14 days, the coin collectors will be about 70% full. The collection agency has planned the visits this way because if the meters are full they become unusable, but emptying the meters too often increases employment costs. During the last visit five randomly selected meters were 50%, 40%, 70%, 75%, and 45% full, respectively. Do you think the frequency of visits should be changed?

mean case, α =0.05
1. μ =average fullness of of meters
μ =1
2. H_0: μ =0.7 H_a: μ !=0.7

n=5,x=0.56 μ =0.7 S=0.15572

t=-2.0103
p=0.115
p> α , fail to reject H_0
Conclusion:
There is not enough evidence in these data to conclude that the frequency of visits should be changed

Poll Question 12.7

Recall: So far, we have discussed strategies for computing confidence intervals and conducting hypothesis tests for a single population parameter μ or p. In many applications it is of interest to compare the parameters of two (or more) populations with each other. There are tests to compare to population means and different tests to compare two population proportions. If you move on to Math 161B you will see both versions in more detail. To give you a taste of more relevant statistics, we'll take a peek at a two-sample procedure to compare two means below.

Inference For Two Sample Means

Example:

- Does a cold last longer/shorter if you take medication?
- Are men better at math than women?

ASSUMPTIONS: Let $X_1, \ldots, X_m$ be a random sample from a population with mean μ_1 and variance σ_1^2. Let $Y_1, \ldots, Y_n$ be a random sample from another population with mean μ_2 and variance σ_2^2. Assume that the samples X and Y are independent.

PROPOSITION: The expected value of $\bar{X} - \bar{Y}$ is $\mu_1 - \mu_2$ with standard deviation

$$\sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

Tests Statistics

CASE 1: If both population distributions are Normal and both standard deviations are known, then the test statistic

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim \text{ Normal}(0, 1)$$

has a Normal distribution.

CASE 2: If both sample sizes are large ($m > 40$ and $n > 40$), then the test statistic

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \sim \text{ Normal}(0, 1)$$

has approximately a Normal distribution.

CASE 3: If the population distributions are Normal but the samples are not very large and the standard deviations are not known then the test statistic

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \sim t_\nu$$

has approximately a t-distribution with df ν estimated by

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)}{\frac{(s_1^2/m)}{m-1} + \frac{(s_2^2/n)}{n-1}}$$

(round ν down to the nearest integer). In a pinch, one can also use the approximation $\nu = \min(n, m) - 1$. The test statistics for the difference in population means can be used to conduct the usual 5-step hypothesis test to decide whether

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$

against one of the alternatives

$$H_a : \mu_1 - \mu_2 > \Delta_0 \quad \text{or} \quad H_a : \mu_1 - \mu_2 < \Delta_0 \quad \text{or} \quad H_a : \mu_1 - \mu_2 \neq \Delta_0$$

But the same information can also be used to derive a $(1-\alpha)\%$ confidence interval for the difference in means.

Recall: Most confidence intervals (and all you learn in this course) have the form

$$\text{CI}_\theta = \left[\hat{\theta} \pm \text{quantile} \cdot \text{SE}_\theta\right]$$

Depending on which case a problem falls into, the distribution of $\bar{X} - \bar{Y}$ is either Normal or t. The standard error is the quantity in the denominator of the test statistic.

Example: The firmness of a piece of fruit is an important indicator of fruit ripeness. The Magness-Taylor firmness (N) was determined for one sample of 20 golden apples after 0 days of shelf life resulting in a sample mean of 8.74 and a sample standard deviation of 0.66. Another sample of twenty of the same kind of apples was tested after 20 days of the shelf with a sample mean of 4.96 and standard deviation 0.39. Calculate a confidence interval for the true difference in firmness that occurs after twenty days on the shelf. Use a confidence level of 95% and interpret the interval.