Day 2, Evening: Univariate Tests of Statistical Inference

Instructor: Marshall A. Taylor
844 Flanner Hall
mtaylo15@nd.edu
marshalltaylor.net

Recap

- This morning we talked about how the asymptotic theory of probability distributions allows us to assess how representative our sample distribution is of the population distribution.

- We specifically looked at how we can quantify the standard deviation of the sampling distribution of sample means (the standard error), which is smaller when we have larger samples sizes. The smaller the standard error, the more reliable our estimates—assuming the sampling distribution is normal per the CLT.

Recap

- We then used the standard error and our assumptions of the sampling distribution to create confidence intervals around the population mean.

- Finally, we talked about z-scores—which transform an observation's value on x so we can see how many standard deviations it is from the variable mean.

But what about random chance?

- Standard errors and confidence intervals around the mean give us a generally idea of how efficient our sample mean might be.

- But how de we *really* know whether or not our sample mean is due to sampling error?

- For example, if our sample mean is different from a known population mean, how do we know whether or not this difference is potentially real or just a product of random chance?

Hypothesis Testing

- Answering a question such as this necessitates **hypothesis testing**.

- A **hypothesis** is a statement about the expected property of a variable or the relationship between multiple variables.

- We **test** a hypothesis by comparing these expectations to observed data to determine whether or not these expected properties or relationships are generalizable to the target population.

Hypothesis Testing

- Such a test first requires stating both a **null hypothesis** and an **alternative hypothesis.**

- A **null hypothesis** (denoted H_0) is a statement that is tested to determine whether its premise should be **accepted** or **rejected**. In statistics, H_0 is typically that there is no statistically significant difference between two numbers.

Hypothesis Testing

- Such a test first requires stating both a **null hypothesis** and an **alternative hypothesis.**

- An **alternative hypothesis** (denoted $\mathrm{H_A}$) is the hypothesis that we **either find support for or not** depending on whether $\mathrm{H_0}$ is rejected or accepted. The alternative hypothesis is often as simple as the proposition that there is a statistically significant difference between two numbers.

Hypothesis Testing

- For example, if we are curious if the hypothesized mean height of ND undergrads (x) finds support given a mean from a random sample of ND undergrads, then our null hypothesis would be:

$$\mathrm{H}_0 : \mu = x$$

Hypothesis Testing

- If we think that the population average is actually taller than the one previously hypothesized, then our alternative hypothesis would be:

$$\mathrm{H_A} : \mu > x$$

- If we think it is shorter, then $\mathrm{H_A}$ would be:

$$\mathrm{H_A} : \mu < x$$

Hypothesis Testing

- Or we might be satisfied with the $\mathrm{H_A}$ that the population average is simply different form the one hypothesized—regardless of the particular direction of the difference. In this case, $\mathrm{H_A}$ would be:

$$\mathrm{H_A} : \mu \neq x$$

Hypothesis Testing

- In hypothesis testing, we test H_0—**not** H_A!
  - If H_0 is accepted, then we do not find support for H_A as something that is happening in the population.
  - If H_0 is rejected, then we do find support for H_A as something that is happening in the population.

Hypothesis Testing

- So we understand this H_0 and H_A business. But how do we go about actually testing H_0?

- For example, how do we know whether or not the population height is significantly different from the one hypothesized?

Back to the theory!

- Luckily, we have the CLT and those critical values to help us out again!

- We can use the CLT to help us assess how likely we are to observe H_A assuming that H_0 is true. If we are not very like to observe H_A when assuming H_0 is true, but we observe it anyway, then we reject the premise of H_0 and instead find support for H_A as something that may be happening in the population.

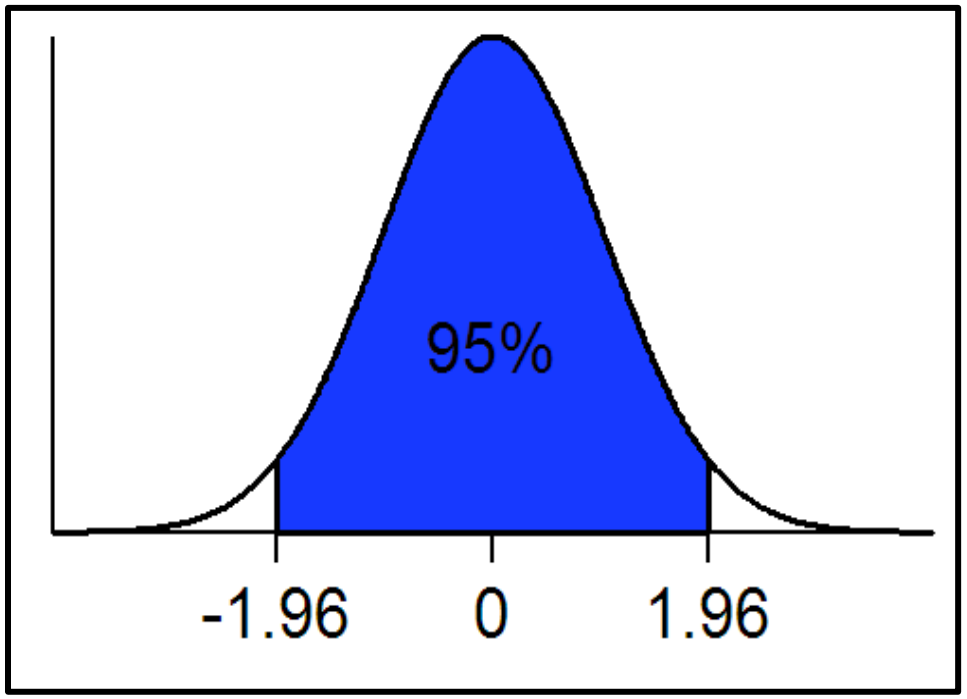
Back to the theory!

Though we have talked about the sampling distribution of sample means up to this point, the CLT actually holds for *any* random variable—standard deviations, differences, whatever!

Back to the theory!

- If we want to know whether or not a sample mean is different from a population mean, then we are interested in the distribution of **mean differences**.

So, if we know that 95% of sample mean differences fall within ±1.96 standard deviations of the "real" mean difference (which we will assume to be 0), then, by way of the complements rule, we also know that 5% do not. So if our observed sample mean difference falls above 1.96 or below –1.96, then we reject H_0 because there is less than a 5% chance that we would observe our sample mean difference if H_0 were true.



An Example: The One Sample *t*-test

- Think back to our height example. Let's say the mean height in our sample is 67 inches (about 5'6"). Now let's hypothesize the mean height of population to be 63 inches (about 5'4"). There are eleven of you, and let's say the sample standard deviation is 3 inches.

- Our H_0 would be that **there is no difference** between the two means ($\mathrm{H}_0 : \mu = 63$).

- Our H_A would be that the population height is taller than the one hypothesized ($\mathrm{H}_\mathrm{A} : \ \mu > 63$).

An Example: The One Sample *t*-test

- We only have one sample statistic that we are working with: the mean height in the sample. As such, we can test our null hypothesis using what is referred to as a **one sample *t*-test**.

- Further, since we are hypothesizing that the difference is in a particular direction, we say that this is a **one-tailed test**. If we were not hypothesizing a particular direction for the difference, this would be a **two-tailed test**.

An Example: The One Sample *t*-test

- The one sample *t*-test directly assesses our H_0 that there is no statistically significant difference by taking the difference between the sample and population mean and scaling this value by the standard error:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

- The result is what is referred to as the ***t* statistic**.

An Example: The One Sample t-test

- This statistic is our **test statistic**. But what do we do with it once we've got it?

An Example: The One Sample *t*-test

- We have to compare our test statistic (a *t*-statistic, in this case) to the critical value associated with our desired confidence level in what is called the ***t*-distribution** (which is very similar to a normal distribution, but "penalizes" you for having small sample sizes and therefore fewer degrees of freedom).

- So let's say that we only want to have a .05 probability of being wrong. The critical *t*-value associated with this (one-tailed) probability, given that we have 10 degrees of freedom (11 – 1), is **1.81**.

- BTW: This ".05 probability of being wrong" is known as our **significance level**, also known as the **alpha level (α).**

An Example: The One Sample *t*-test

If our test statistic falls above 1.81, then we can reject $\mathbf{H_0}$ because our statistic does not fall within the bottom 95% of the normal distribution if we assume the "real" mean difference is 0!

An Example: The One Sample *t*-test

- If we could reject H_0, then our test statistic (the *t* statistic, in this case) would fall into the **right tail of distribution**, as shown below (with the difference being that the critical value is 1.64 as opposed to 1.81).

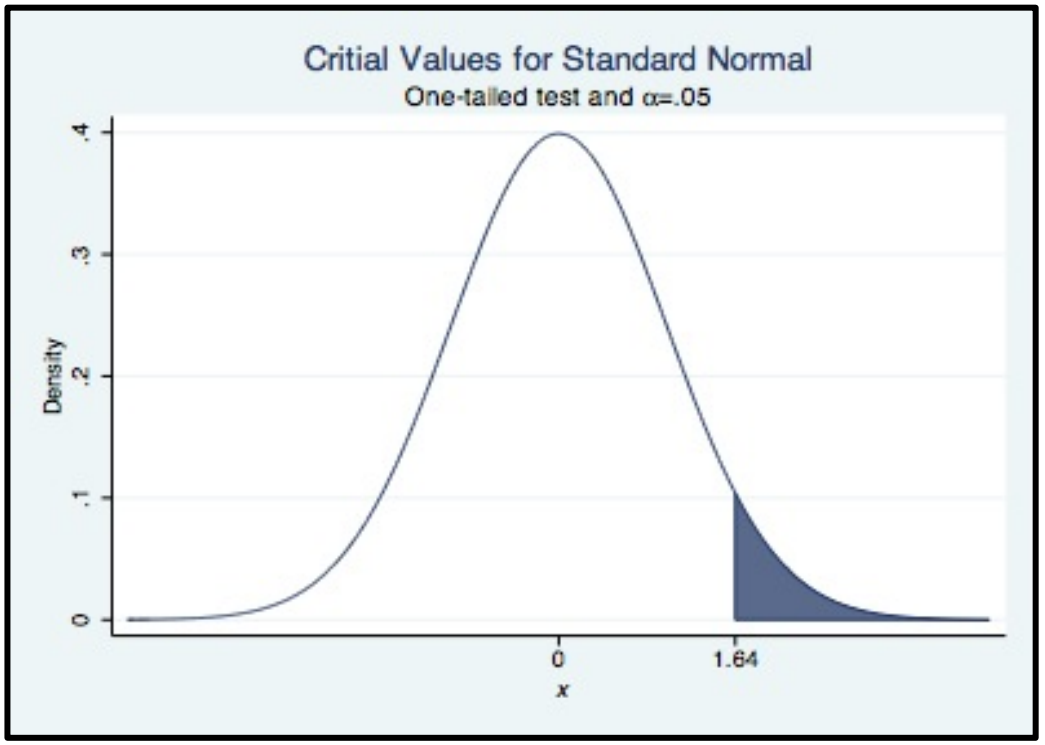


*Image from PsychStatistics website (http://www.psychstatistics.com/2010/11/24/stata-graphing-distributions/).

An Example: The One Sample *t*-test

- So let's crunch the numbers for our height example:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{67 - 63}{\frac{3}{\sqrt{11}}} = 4.42$$

An Example: The One Sample *t*-test

- Our test statistic is **4.42**. This is much larger than **1.81**.

- As such, we can reject the null hypothesis—**at the .05 level**—that the hypothesized population mean height is 63 inches. Maybe we should therefore reconsider the "real" value of the population mean (i.e., all ND undergrads).

- If we are absolutely sure that the population mean height is 63 inches, then we can at least say that it is very unlikely that our sample mean came from a population with a mean of 63. Perhaps our sample is not entirely random, or maybe we just got a bad sample.

An Example: The One Sample *t*-test

- Confirm with Stata:

```
. ttesti 11 67 3 63, level(95)

One-sample t test

------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      11          67    .904534           3    64.98457    69.01543
------------------------------------------------------------------------------
    mean = mean(x)                                t =   4.4222
Ho: mean = 63                     degrees of freedom =       10

    Ha: mean < 63                 Ha: mean != 63                 Ha: mean > 63
 Pr(T < t) = 0.9994         Pr(|T| > |t|) = 0.0013          Pr(T > t) = 0.0006
```

Let's try another one

- A random sample of five congressional candidates from 2006 raised, on average, $1,964,018. Now let's say that the average funds raised across all congressional candidates in the country in 2006 was hypothesized to be $5,000,000.* Further, the sample candidates vary from their mean of $1,964,018 by about $925,149. At $\alpha = .05$, how likely is it that our hypothesized mean value reflects the population parameter (regardless of direction)?

- Tip: the critical t-value with four degrees of freedom (5 candidates – 1) is ± 2.776.

*Made up number.

Let's try another one

- First state your null hypothesis:

$$H_0 : \mu = \$5{,}000{,}000$$

- Then the alternative hypothesis that they are different:

$$H_A : \mu \neq \$5{,}000{,}000$$

- This is a **two-tailed test**, since we are not assuming a particular direction of the difference. Thus the $\neq$ rather than the $<$ or $>$!

Let's try another one

- Specifically, we **reject** H_0 if our test statistic falls **above** 2.776 or **below** –2.776, as illustrated by the shaded regions in the distribution below:

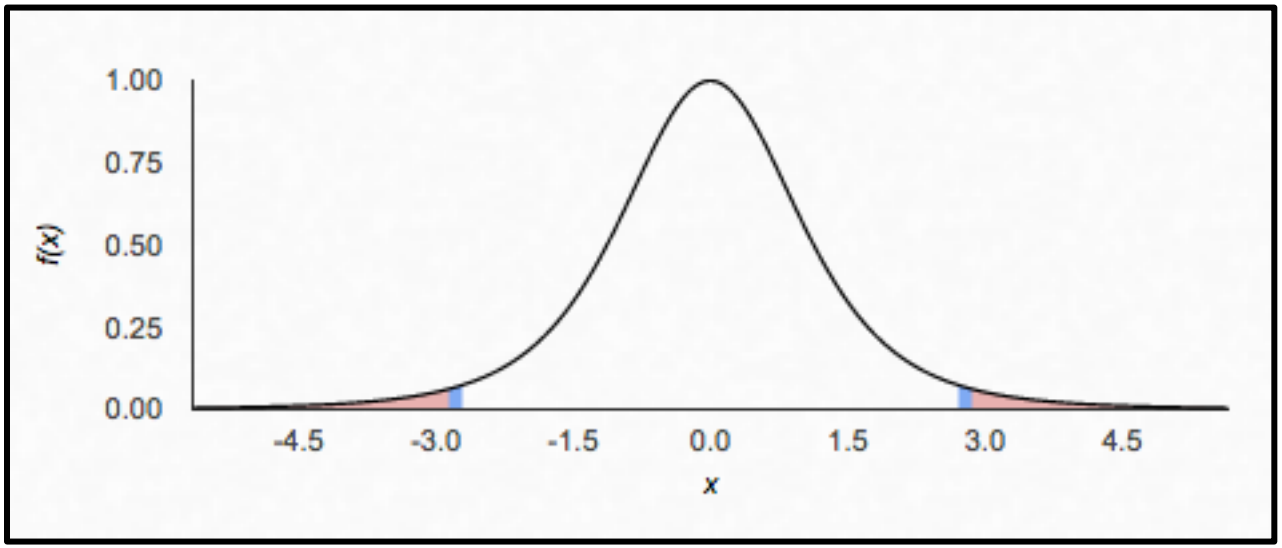


Figure created using Student's t-distribution applet at the University of Iowa (http://homepage.divms.uiowa.edu/~mbognar/applets/t.html).

Let's try another one

- Remember our formula: $t = (\bar{x} - \mu)/(s/\sqrt{n})$

- So:

$$t = \frac{1964018 - 5000000}{(\frac{925149}{\sqrt{5}})} = -7.338$$

Let's try another one

- Our test statistic is **–7.338**. This is smaller than our critical value, –2.776.
- So, at the .05 significance level, we **reject the null hypothesis** that the population mean funds raised is $5,000,000.
- Instead, we **find statistical support for our alternative hypothesis** that the population mean funds raised is **not** $5,000,000.
- If we are sure that the population mean is $5,000,000 (i.e., it is more than a *hypothesized* mean), then we can at least say that it is very unlikely that our sample mean could have come from a population with a mean of $5,000,000.

Let's try another one

- Now let's say we wanted to know if the difference is significant in a particular direction. Specifically, we want to know if the population mean is smaller than the one hypothesized.

- In this case: $H_A : \mu < \$5{,}000{,}000$

Let's try another one

- Our new critical value is **–2.132**. This value is smaller than the one from our two-tailed test, meaning that we have more **power** to reject the H_0 (because the shaded region is bigger). **However**, it comes at the expense of not being able to detect significant test statistics at the other end of the distribution.

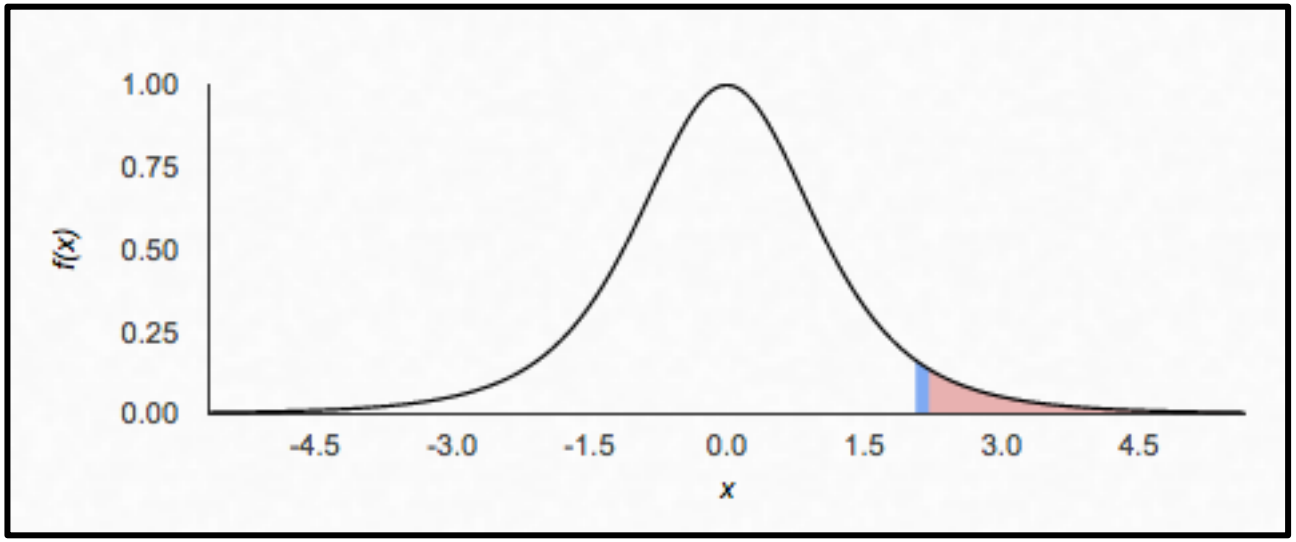


Figure created using Student's *t*-distribution applet at the University of Iowa (http://homepage.divms.uiowa.edu/~mbognar/applets/t.html).

Let's try another one

- So –7.338 is smaller than our critical value of –2.132.
- We can once again reject the H_0.
- However, we can now say that it is likely (at the .05 significance level) that the real population means is **less** than the hypothesized population mean.

But let Stata do the work!

- Luckily, Stata can handle this.

```
. ttest FRAISED==5000000 if Y2006==1 & STATE2==14

One-sample t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 FRAISED |       5     1964018    413739.2      925149    815293.9     3112742
------------------------------------------------------------------------------
    mean = mean(FRAISED)                                          t =  -7.3379
Ho: mean = 5000000                              degrees of freedom =        4

 Ha: mean < 5000000           Ha: mean != 5000000           Ha: mean > 5000000
 Pr(T < t) = 0.0009         Pr(|T| > |t|) = 0.0018          Pr(T > t) = 0.9991
```

But let Stata do the work!

- Luckily, Stata can handle this.

```
. ttest FRAISED==5000000 if Y2006==1 & STATE2==14

One-sample t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 FRAISED |       5     1964018    413739.2      925149    815293.9     3112742
------------------------------------------------------------------------------
    mean = mean(FRAISED)                                          t =  -7.3379
Ho: mean = 5000000                              degrees of freedom =        4

 Ha: mean < 5000000           Ha: mean != 5000000           Ha: mean > 5000000
 Pr(T < t) = 0.0009         Pr(|T| > |t|) = 0.0018          Pr(T > t) = 0.9991
```

We got our test statistic correct.

But let Stata do the work!

- Luckily, Stata can handle this.

```
. ttest FRAISED==5000000 if Y2006==1 & STATE2==14

One-sample t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 FRAISED |       5     1964018    413739.2      925149    815293.9     3112742
------------------------------------------------------------------------------
    mean = mean(FRAISED)                                          t =  -7.3379
Ho: mean = 5000000                              degrees of freedom =        4

 Ha: mean < 5000000           Ha: mean != 5000000          Ha: mean > 5000000
 Pr(T < t) = 0.0009         Pr(|T| > |t|) = 0.0018         Pr(T > t) = 0.9991
```

And both our two-tailed and one-tailed tests were correct. The *p*-value for the one-tailed test tells us that there is only a .09% chance that our sample mean would differ by at least this much from the population when we assume that the "real" mean difference is zero.

But let Stata do the work!

- Luckily, Stata can handle this.

```
. ttest FRAISED==5000000 if Y2006==1 & STATE2==14

One-sample t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 FRAISED |       5     1964018    413739.2      925149    815293.9     3112742
------------------------------------------------------------------------------
    mean = mean(FRAISED)                                          t =  -7.3379
Ho: mean = 5000000                              degrees of freedom =        4

 Ha: mean < 5000000           Ha: mean != 5000000          Ha: mean > 5000000
 Pr(T < t) = 0.0009         Pr(|T| > |t|) = 0.0018         Pr(T > t) = 0.9991
```

But notice our one-tailed test in the other direction would have given us a non-significant result.

Other tests

- But sometimes we need to test statistics other than means. Remember, means aren't always an appropriate measure of central tendency!

- Sometimes we have a variable that can only take on two categories. In this case, we can use the **one sample difference-of-proportions test**.

- Other times the median is a better measure of central tendency. In this case, we can use the **one sample sign test**.

One Sample Difference-of-Proportions Test

- Very similar to the one sample *t*-test, but for a dichotomous variable instead of a mean.

- The test statistic can be found with:

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

- Notice the test statistic is a *z*-statistic instead of a *t*-statistic. This simply means that we are assuming a sampling distribution that follows a normal distribution instead of a *t*-distribution (which, with a large enough sample, converge to essentially be same thing).

An Example: The One Sample Difference-of-Proportions Test

- A landlord at a busy downtown apartment building took a poll and found that only 3% of residents are in favor of requiring parking passes. You question the landlord's conclusion and take a random sample of 50 residents. You find that, among these residents, 23 are in favor of requiring parking passes—46% of your sample. How valid is the landlord's conclusion, assuming that you are only willing to be wrong 5% of the time?

An Example: The One Sample Difference-of-Proportions Test

- Our null hypothesis:

$$H_0 : \mu = .03$$

- If we take on a one-tailed test, our alternative hypothesis would be:

$$H_A : \mu > .03$$

.

An Example: The One Sample Difference-of-Proportions Test

- Our critical z-value is **1.645**. We choose this value instead of ± 1.96 because we are performing a one-tailed test. So, instead of rejecting H_0 if our test statistic falls within the top and bottom 2.5% of the normal distribution (bottom left plot), we look at the top 5% only (bottom right plot).

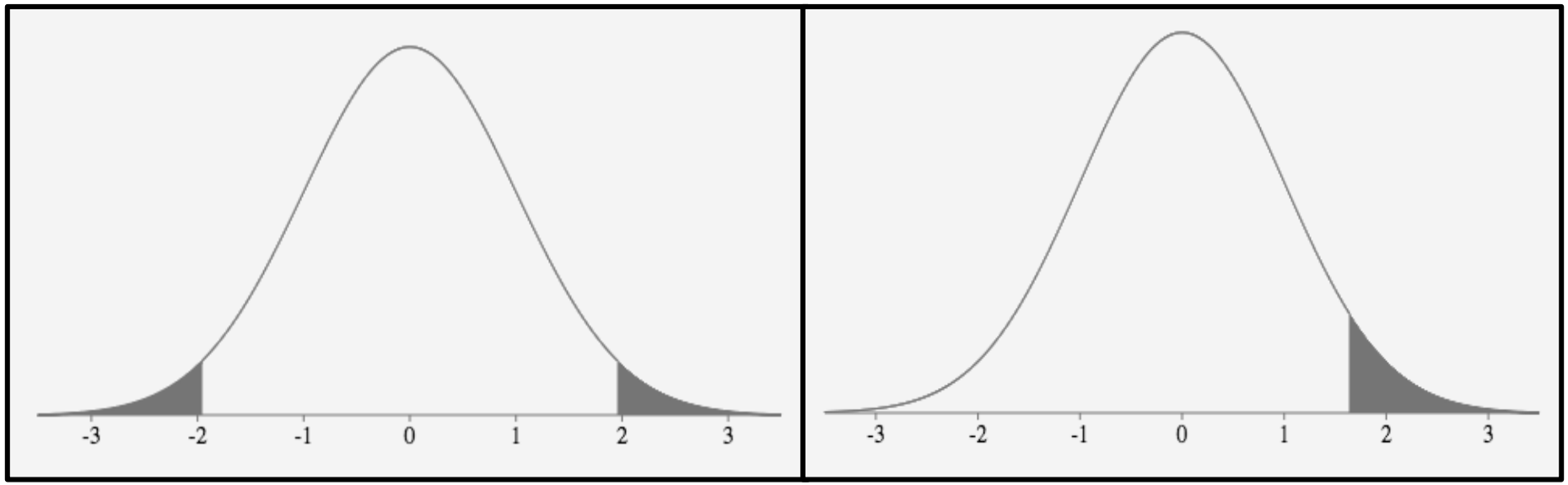


*Figure made with Normal Distribution Calculator at Online Statsbook (http://onlinestatbook.com/2/calculators/normal_dist.html).

An Example: The One Sample Difference-of-Proportions Test

- Now let's do the math:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.46 - .03}{\sqrt{\frac{.03(1-.03)}{50}}} = 6.632$$

- Our test statistic is much larger than our critical value of 1.645. As such, there is less than a 5% chance that we would find 46% of 50 residents in favor of parking passes if it were true that only 3% of total residents were in favor of parking passes. We conclude that more residents are in favor of this policy at the .05 significance level.

.

A Stata Example

- We know that 52% of our sample of House candidates in 2006 “went negative”—that is, they engaged in negative campaigning against their opponent. Let’s say (hypothetically) that previous studies suggest that, on average and across time, House candidates go negative about 55% of the time. If our sample consists of 227 candidates and is indeed a representative random sample of House candidates across time, can we say that House candidates actually go negative less than 55% of the time?

A Stata Example

```
. prtest GONEG==.55 if Y2006==1 & SENATE==0

One-sample test of proportion                 GONEG: Number of obs =      227
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.                      [95% Conf. Interval]
-------------+----------------------------------------------------------------
       GONEG |   .5242291   .0331472                      .4592618    .5891964
------------------------------------------------------------------------------
    p = proportion(GONEG)                                         z =  -0.7805
Ho: p = 0.55

   Ha: p < 0.55                 Ha: p != 0.55                 Ha: p > 0.55
 Pr(Z < z) = 0.2176         Pr(|Z| > |z|) = 0.4351          Pr(Z > z) = 0.7824
```

- We cannot reject the null hypothesis at the .05 level that House candidates, on average and across time, go negative 55% of the time.

One Sample Sign Test

- Sometimes the median is the most efficient measure of central tendency.
  - Perhaps the variable is ordinal.
  - Or maybe it is a continuous variable but skewed because of outliers.

- In this case our best univariate test of statistical inference is the **one sample sign test**.

One Sample Sign Test

- The test statistic for this test is (1) in the case of one-sided tests, the number of positive or negative differences between the observed values and the hypothesized median value, or (2) in the case of two-sided tests, the smaller of these two differences. In either case, the number of positive (P) differences and negative (N) differences is calculated with $\Sigma x_{\mathrm{i}} - m_0$. They are then sorted based on the sign of the difference.

One Sample Sign Test

- The null hypothesis for this test is that there is no statistically significant difference between the sample median and the hypothesized median ($\mathrm{H}_0 : m = m_0$).

- The two-tailed alternative hypothesis would be:

$$\mathrm{H}_\mathrm{A} : m \neq m_0$$

- With one-tailed tests, the H_A can be one of:

$$\mathrm{H}_\mathrm{A} : m > m_0$$
$$\mathrm{H}_\mathrm{A} : m < m_0$$

One Sample Sign Test

- We then calculate the probability of observing the differences (d) that we do under the assumption that the "real" difference is 0:

$$P(d) = \frac{n!}{d!(n-d)!} p^d (1-p)^{n-d}$$

  - Where n is the sample size minus "ties" and p is the probability of observing either a positive or negative difference when the difference between the sample median and hypothesized median is assumed to be 0—that is, p = .5, since we are just as likely to observe either kind of difference.

An Example: One Sample Sign Test

- Nine students are surveyed to assess how much they enjoy the quality of their school food. They are given a Likert scale that ranges between 1 (very bad) to 5 (very good). The observed scores are 1, 3, 5, 4, 2, 3, 3, 4, and 5. Previous surveys suggest that the median rating for food quality among students is 4 (fairly good). Do we have sufficient evidence to infer that students now rate the food quality more poorly?

An Example: One Sample Sign Test

- First, we find the number of positive and negative differences:

5 negatives

Observed	m_0	Difference
1	4	-3
3	4	-1
5	4	1
4	4	0
2	4	-2
3	4	-1
3	4	-1
4	4	0
5	4	1

2 positives

2 ties (0s)

An Example: One Sample Sign Test

- We are interested if students rate food quality **more poorly**, so this is a one-tailed test and we focus on the negative differences (d = 5):

$$P(d) = \frac{7!}{5!(7-5)!}.5^5(1-.5)^{7-5} = .164$$

- The number .164 represents the probability that we would observe this many negative differences if the medians were equal.

An Example: One Sample Sign Test

- We then repeat this procedure to obtain the probabilities for getting 6, 7, 8, and 9 negative differences.

- After doing this we add up the probabilities: .227.

- This is much larger than our cut-off value of .05, so we cannot reject the null hypothesis that there is no statistically significant difference between our sample median and the hypothesized median. Given these data, we cannot say that students now rate their school food quality any lower than they used to.

Yay for Stata!

```
. signtest var1=4

Sign test

        sign |     observed    expected
-------------+------------------------
    positive |            2         3.5
    negative |            5         3.5
        zero |            2           2
-------------+------------------------
         all |            9           9

One-sided tests:
  Ho: median of var1 - 4 = 0 vs.
  Ha: median of var1 - 4 > 0
      Pr(#positive >= 2) =
         Binomial(n = 7, x >= 2, p = 0.5) =  0.9375

  Ho: median of var1 - 4 = 0 vs.
  Ha: median of var1 - 4 < 0
      Pr(#negative >= 5) =
         Binomial(n = 7, x >= 5, p = 0.5) =  0.2266

Two-sided test:
  Ho: median of var1 - 4 = 0 vs.
  Ha: median of var1 - 4 != 0
      Pr(#positive >= 5 or #negative >= 5) =
         min(1, 2*Binomial(n = 7, x >= 5, p = 0.5)) =  0.4531
```

Conclusion

- Our knowledge of sampling distributions is what allows us to get a grasp on the likelihood that our sample estimates are “generalizable” to the target population or due to sampling error (random chance).

- We make these inferences through hypothesis testing.

- We explored a variety of univariate inferential tests. They are univariate because we compare just one sample estimate to hypothesized population parameters.

- Tomorrow we start looking at bivariate inferential tests; i.e., comparisons between two sample estimates.

Datasets Used

- Druckman, James, Michael Parkin, and Martin Kifer. 2013. *Congressional Candidate Websites*. ICPSR-34895-v1. Ann Arbor, MI: Inter-University Consortium for Political and Social Research. Retrieved February 10, 2015 (http://doi.org/10.3886/ICPSR34895.v1).

- The Stata survey documentation data, `nhanes2f`, from the *Stata Press* website. Retrieved July 24, 2016 (http://www.stata-press.com/data/r11/svy.html).