Chapter 2
Review of Hypothesis Testing and Basic Tests

Review of Basic Concepts:
Review of Hypothesis Testing

1. Logic of Hypothesis Testing

- Because of random variability in the data, the parameters we estimate from the data will never match the population parameters exactly (see 1-48 and 1-49)

- This fact presents a problem for us. Consider the speeding example. Suppose we want to know if, on average, people in this sample are speeding. In our sample of 25, we found the average speed to be 62.5 MPH. Although this number is larger than 55 MPH, there are two reasons why this sample mean could be larger than 55MPH

  - The true value speed in the population is greater than 55 MPH. In other words, we sampled from a distribution that had a mean value of speed greater than 55.
    - If this is the case, then we should conclude that on average this population of drivers violates the speed limit
    - If we repeated the sampling process, it is likely we would again find a sample mean greater than 55MPH

  - A second possibility is that we sampled from a distribution that had a mean value of speed equal to or less than 55, but because of the random variability in the sampling process, we happened to obtain a sample with an average speed of 62.5 MPH.
    - If this is the case the population of drivers does not, on average, violate the speed limit
    - Our findings are not due to sampling from a population of speeders, but are due to random chance
    - If we repeated the sampling process, it is likely we would find a sample mean near 55 MPH

- Hypothesis Testing is the process of performing a statistical test to determine the likelihood that the estimate/association was seen by chance alone

- The statistical hypothesis vs. the research hypothesis

  - The research hypothesis represents the rationale of a study and specifies the kinds of information required to support that hypothesis
  - To evaluate the research hypothesis statistically, the researcher must form a set of statistical hypotheses
    - It is the statistical hypotheses that are assessed and evaluated in the statistical analysis. We use the outcomes of the statistical hypotheses to evaluate and refine the research hypothesis
    - The statistical hypothesis establishes a set of mutually exclusive and exhaustive hypotheses about the true value of the parameter in question

- The null and alternative hypotheses
  - The starting point of any statistical hypothesis is the null hypothesis
  - The null hypothesis is the statement that the observed data do not differ from what would be expected on the basis of chance alone

    - Example 1: You collect speed data to determine if drivers are violating the speed limit

      $$H_0 : \mu_{MPH} \leq 55$$

    - Example 2: You investigate if a GRE prep class improves GRE scores by giving people a GRE test before the class and comparing the result to a GRE after the class

      $$H_0 : \mu_{GRE_{before}} = \mu_{GRE_{after}}$$

      Or more generally,

      $$H_0 : \mu_{Time_1} = \mu_{Time_2}$$

    - Example 3: You design a study to compare two new drugs to a no drug control

      $$H_0 : \mu_{Control} = \mu_{DrugA} = \mu_{DrugB}$$

      Or more generally,

      $$H_0 : \mu_{Group_1} = \ldots = \mu_{Group_m}$$

    - Example 4: You measure self-esteem and depression scores to see if they are related to each other

      $$H_0 : \rho = 0$$

- For each null hypothesis, you must have a corresponding alternative hypothesis. The null and alternative hypotheses must be:
  - Mutually exclusive (no outcome can satisfy both the null and alternative hypotheses)
  - Exhaustive (every possible outcome must satisfy either the null or alternative hypothesis)

  - As a result, once you specify the null hypothesis, the alternative hypothesis is automatically determined

  - Example 1: Speeding example

$$H_0 : \mu_{MPH} \leq 55$$
$$H_1 : \mu_{MPH} > 55$$

  - Example 2: GRE example

$$H_0 : \mu_{GRE_{before}} = \mu_{GRE_{after}}$$
$$H_1 : \mu_{GRE_{before}} \neq \mu_{GRE_{after}}$$

    Or more generally,

$$H_0 : \mu_{Time_1} = \mu_{Time_2}$$
$$H_1 : \mu_{Time_1} \neq \mu_{Time_2}$$

  - Example 3: New drug example

$$H_0 : \mu_{Control} = \mu_{DrugA} = \mu_{DrugB}$$
$$H_1 : \textit{At least one } \mu_i \textit{ differs from the other means}$$

    Or more generally,

$$H_0 : \mu_{Group_1} = \ldots = \mu_{Group_m}$$
$$H_1 : \textit{At least one } \mu_i \textit{ differs from the other means}$$

  - Example 4: Self-esteem and depression

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

- Be sure to always state the null and alternative hypotheses in terms of the population parameters, μ, σ, ρ, and NOT the sample statistics, $\overline{X}, s, r$

- The Logic of the Null Hypothesis Test

  - All hypothesis testing starts with the assumption that the null hypothesis is true. It is the hypothesis of no association, or that the groups came from the same underlying distribution.

  - There are three possible explanations for any differences that might be observed:
    1. An all systematic factors explanation
    2. An all chance explanation
    3. A chance + systematic factors explanation

    - But an all systematic explanation never happens
    - We are left to choose between a chance explanation and a chance + systematic factors explanation
    - The null hypothesis test examines the viability of the all chance explanation.
      ⇒ If differences in the data can be accounted for by a random process, we fail to reject the null hypothesis (Note that we are not saying that the null hypothesis is correct – just that we do not have enough evidence to eliminate this possibility)
      ⇒ If an all chance explanation cannot explain the data, then we reject the null hypothesis, and we conclude that there must be some systematic factor at work.

    - The statistical hypothesis test gives us no information as to what the systematic factor might be. It could be the experimenter's manipulation, or a naturally occurring association/difference between the groups. It is up to the researcher to identify a plausible systematic factor. It is at this point that we must relate the statistical hypothesis to our research hypothesis.

    - See Abelson's (1995) ESP example

- How confident do we have to be that the null hypothesis is false before we reject it?
  - Suppose a shady fellow gives you a coin, and you want to test if the coin is biased toward heads
  - We can convert this research hypothesis to a statistical hypothesis

$$H_0 : \Pi_{Heads} \leq .5$$
$$H_1 : \Pi_{Heads} > .5$$

  - Suppose we flip the coin one time and get a heads. Are we confident that the coin is biased toward heads?

    - No! If the null hypothesis is true, then the coin had a 50% probability of coming up heads. If we rejected the null hypothesis then when the null hypothesis is correct, we would be wrong 50% of the time!

  - Suppose we flip the coin two times and get two heads. Are we confident that the coin is biased toward heads?

    - We can use the binomial theorem to determine the probability of observing two heads in two coin tosses

$$p(x) = \binom{N}{x} p^x 1 - p^{(N-x)}$$

Where N = total number of trials
x = # of successes
p = probability of success

$p(0;2;.5) = .25$
$p(1;2;.5) = .50$
$p(2;2;.5) = .25$

    - If the null hypothesis is true, then the coin had a 25% probability of coming up heads on both tosses. If we rejected the null hypothesis then when the null hypothesis is correct, we would be wrong 25% of the time!

- Suppose we flip the coin three times and get three heads. Are we confident that the coin is biased toward heads?

  $p(0;3;.5) = .125$
  $p(1;3;.5) = .375$
  $p(2;3;.5) = .375$
  $p(3;3;.5) = .125$

  - If the null hypothesis is true, then the coin had a 12.5% chance of coming up heads on all three tosses. If we rejected the null hypothesis then when the null hypothesis is correct, we would be wrong 12.5% of the time. I'd still feel a bit uneasy about rejecting the null hypothesis in this case

- Suppose we flip the coin four times and get four heads. Are we confident that the coin is biased toward heads?

  $p(0;4;.5) = .0625$
  $p(1;4;.5) = .25$
  $p(2;4;.5) = .375$
  $p(3;4;.5) = .25$
  $p(4;4;.5) = .0625$

  - If the null hypothesis is true, then the coin had a 6.25% chance of coming up heads on all four tosses. If we rejected the null hypothesis then when the null hypothesis is correct, we would be wrong 6.25% of the time.
  - Perhaps that might be good enough for us to conclude that the coin is biased. However, the scientific convention has been that when the null hypothesis is true you need to have a probability of .05 or less so that the observed result (or a more extreme result) could be due to chance alone

- Suppose we flip the coin nine times. How many heads would we have to observe to be confident that the coin is biased toward heads?

$p(0;9;.5) = .0020$	$p(5;9;.5) = .2461$
$p(1;9;.5) = .0176$	$p(6;9;.5) = .1641$
$p(2;9;.5) = .0703$	$p(7;9;.5) = .0703$
$p(3;9;.5) = .1641$	$p(8;9;.5) = .0176$
$p(4;9;.5) = .2461$	$p(9;9;.5) = .0020$

  - If the null hypothesis is true (the coin is fair), then the probability of observing 8 or 9 heads in 9 coin flips is:
    $p(x = 8 \text{ or } x = 9) = .0176 + .0020 = .0195$

  - If the null hypothesis is true (the coin is fair), then the probability of observing 7 or more heads in 9 coin flips is:
    $p(x = 7 \text{ or } x = 8 \text{ or } x = 9) = .0703 + .0176 + .0020 = .0898$

  - Thus, we would need to observe 8 or more heads (out of nine tosses) to reject the null hypothesis and conclude that the coin is biased toward heads

- Some terminology regarding hypothesis testing:

  - α (the alpha level or the significance level)
    - The probability of rejecting the null hypothesis when the null hypothesis is true
    - Also known as a Type I error
    - By convention, usually $\alpha = .05$

  - *p*-value (or probability value)
    - The probability of observing an outcome as extreme as (or more extreme than) the observed value, if the null hypothesis is true
    - If $p \le \alpha$ then we reject the null hypothesis
    - If $p > \alpha$ then we retain the null hypothesis
      (or we fail to reject the null hypothesis)

- A one-tailed vs. a two tailed hypothesis test
  - A one-tailed hypothesis test specifies a direction of the effect:

$$H_0 : \Pi_{Heads} \leq .5$$
$$H_1 : \Pi_{Heads} > .5$$

  - A two-tailed hypothesis test is non-directional:

$$H_0 : \Pi_{Heads} = .5$$
$$H_1 : \Pi_{Heads} \neq .5$$

  - Our coin-tossing example was an example of a one-tailed test. We only examined one tail of the distribution (the possibility that the coin was biased toward heads)

  - We could have tested a two-tailed hypothesis (the possibility that the coin was biased):

$p(0;9;.5) = .0020$ $p(5;9;.5) = .2461$
$p(1;9;.5) = .0176$ $p(6;9;.5) = .1641$
$p(2;9;.5) = .0703$ $p(7;9;.5) = .0703$
$p(3;9;.5) = .1641$ $p(8;9;.5) = .0176$
$p(4;9;.5) = .2461$ $p(9;9;.5) = .0020$

    For a one tailed test, we only looked at the biased toward heads end of the distribution:

$p(x \geq 8) = .0195$
$p(x \geq 7) = .0898$

    For a two tailed test, we would also need to consider the possibility that the coin was biased toward tails:

$p(0 \leq x \text{ and } x \geq 9) = .0039$
$p(1 \leq x \text{ and } x \geq 8) = .0390$
$p(2 \leq x \text{ and } x \geq 7) = .1797$

  - In science, there is a very strong preference toward two-tailed tests. It is important that you know how to conduct and interpret one-tailed tests, but in practice most statistical tests will be two-tailed.

- The general procedure of the null hypothesis test:
  - State the null and alternative hypotheses
  - Specify α and the sample size
  - Select an appropriate statistical test

    (Note that all of the preceding steps should be conducted BEFORE collecting data!)

  - Compute the test statistic based on the sample data
  - Determine the *p*-value associated with the statistic
  - Make the decision by comparing the *p*-value to α
  - Report your results (*ALWAYS* including effect sizes)

- Types of errors in hypothesis testing

State of the world	Decision: Reject Null Hypothesis	Decision: Fail to Reject Null Hypothesis
Null Hypothesis TRUE	Type 1 Error Probability = α	Correct Decision Probability = 1 - α
Null Hypothesis FALSE	Correct Decision Probability = 1 – β (POWER)	Type 2 Error Probability = β

  - Interpretation of the *p* value – What does it mean to say *p* = .04? Which statement is true?

    i. The probability that there is no difference between the groups is .04
    ii. Assuming that the null hypothesis is true, the probability we would have observed a difference this large (or larger) is .04

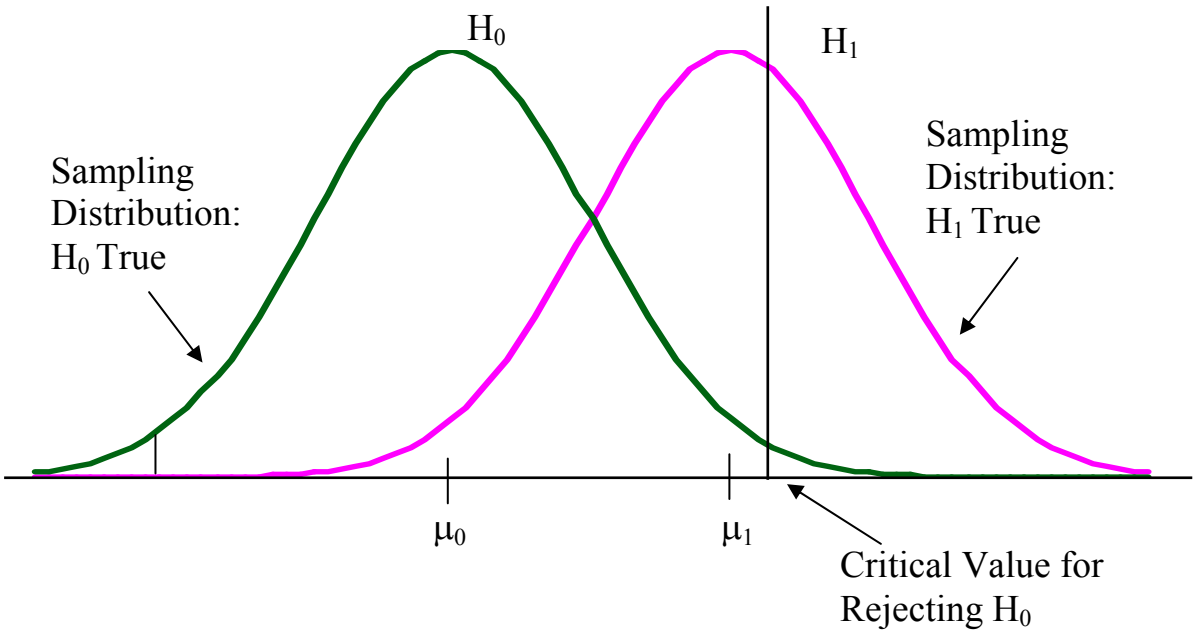
    - Option 1 is P(H_0 | rejection)
    - Option 2 is P(rejection | H_0)

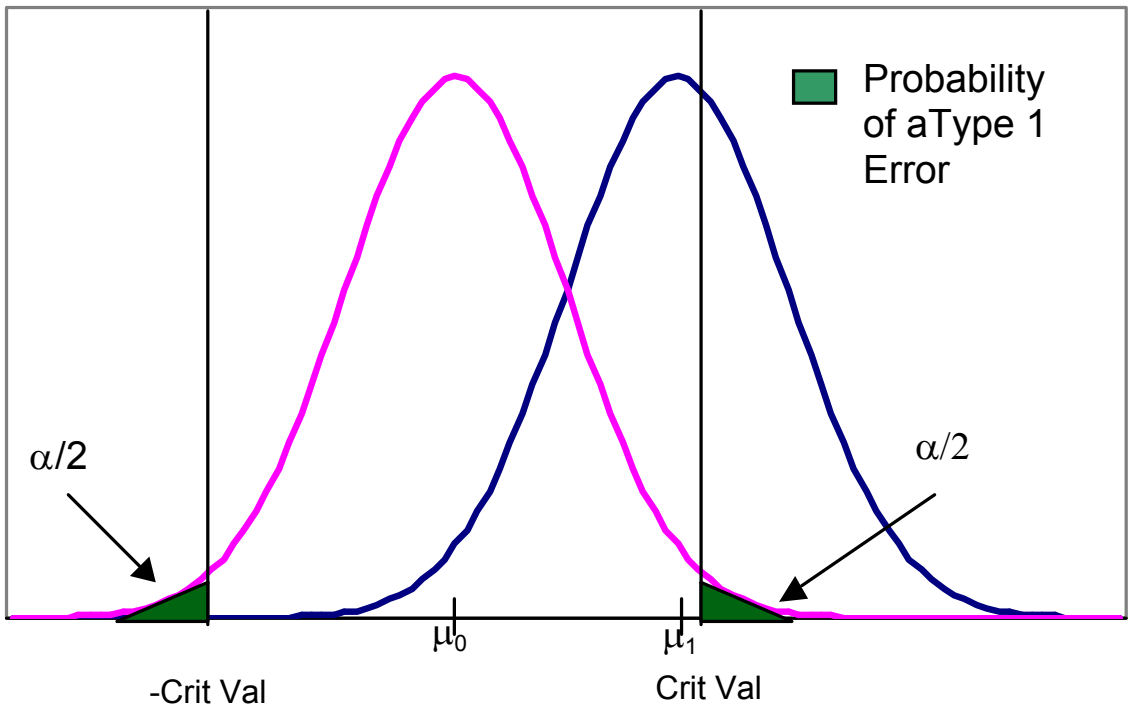
    - Option 2 is the correct statement (The *p*-value is calculated under the assumption that the null hypothesis is true). Do not get confused!

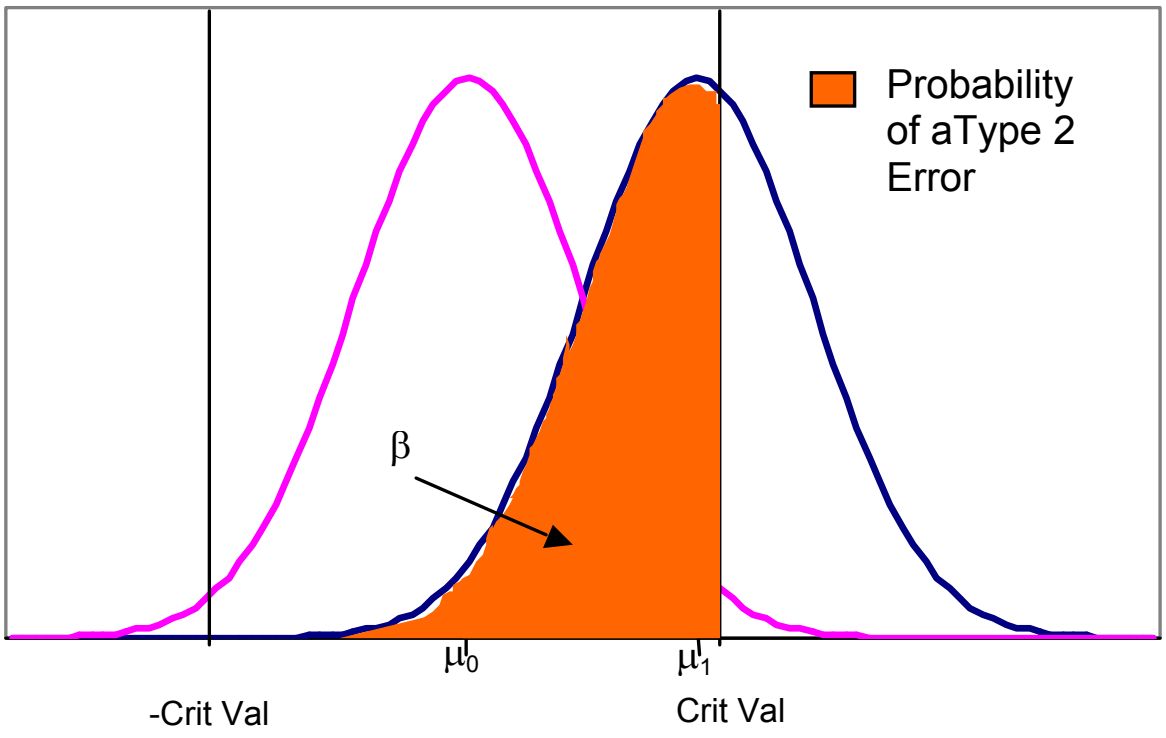
    - For an excellent discussion of issues regarding interpretation of p-values, and the use of confidence intervals and effect sizes, see:

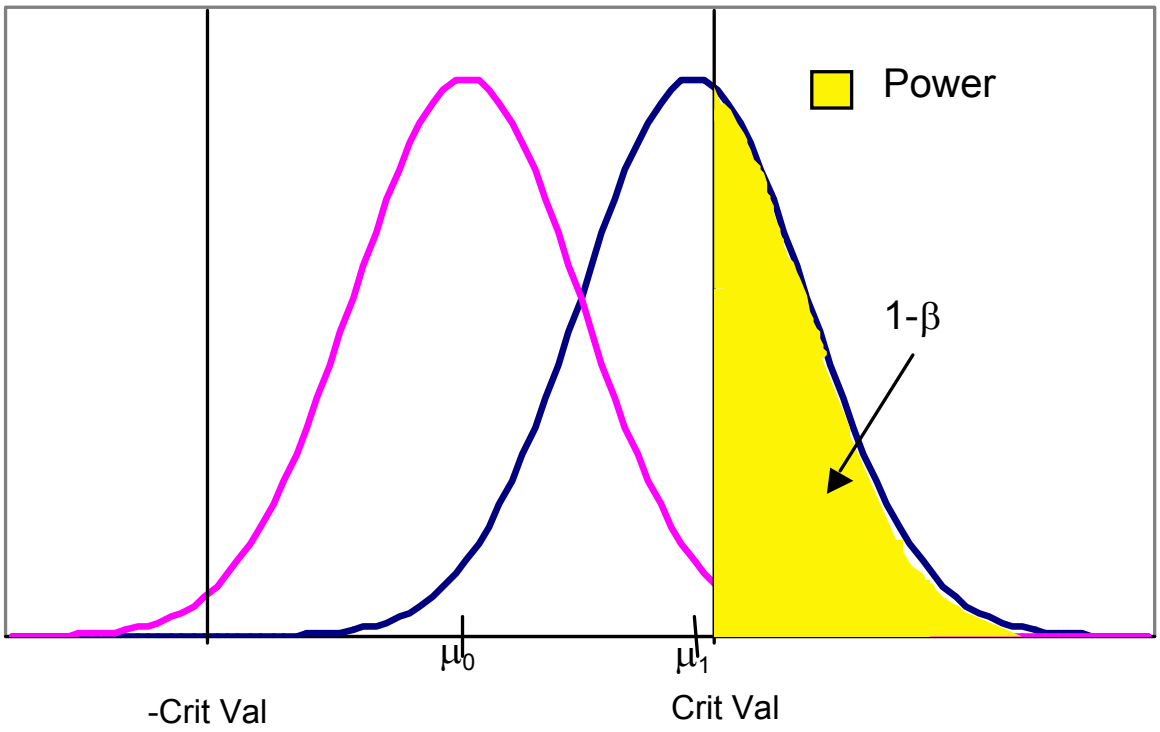
      Cohen, J. (1994). The earth is round (*p* < .05). *American Psychologist, 49*, 997-1003.

- Determining α and 1-β from the sampling distributions:









- The power of a test can be increased by
  - Increasing the distance between μ_0 and μ_1 (effect size)
  - Decreasing the standard deviation of the sampling distributions (Usually by increasing the sample size, but also by decreasing σ)
  - Increasing the probability of a Type 1 error (α)

Significance test $\propto$ effect size * sample size (N)

- Of course, knowing that you had low power after you ran the test does little good . . .

2. One sample z-test
   - In our previous discussion of the z-scores, we examine probability (or percentile) of an individual score.
     - But we may also want to know if a mean from a sample drawn from a known population differs from that population mean?
     - The test used to answer this question is known as the <u>one sample z-test of population mean difference</u>

   - Note that in this question, the sample is drawn from a known population. In other words, before the data were collected, we already knew the distribution of the population. In order to use a z-test:
     - The population must be normally distributed
     - The population parameters must be known in advance of the study
     - The observations must be independent and randomly drawn from the population

   - In general, any test statistic will have the form:

$$test_{obs} = \frac{observed - expected}{standard\ error}$$

     - In this case, we are testing if an observed sample mean is equal to a known population value. Thus, the observed value is the observed sample mean, $\overline{X}$, and the expected value is the known population value, μ

     - For the denominator of the test, we need to calculate the standard error of the observed sample mean. But we have already done so! We know that the estimated standard error of a sample mean is

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}}$$

     - Putting both pieces together, we arrive at the test statistic for the one-sample z-test:

$$z_{obs} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{N}}}$$

- Example 1: National norms for a high school proficiency test are distributed $N(75,16)$. A random sample of 120 high school seniors has a mean proficiency score of 72.5 ($\overline{X}_{obs} = 72.5$). Do these sample scores differ significantly from the overall population mean (use $\alpha = .05$)?

  - In advance, we know the population has a $N(75,16)$ distribution. We wish to compare the mean of one sample to the known population mean, and the observations are independent. Thus, the one sample z-test is appropriate.

- Example 1a: First, let's conduct a one tailed test that the sample of high-school seniors have a lower score than the national average.

  - State Null and Alternative Hypotheses

    $H_0 : \mu \geq 75$

    $H_1 : \mu < 75$

  - Specify α and the sample size

    $\alpha = .05 \qquad N = 120$

  - Compute the test statistic z_{obs} based on the sample data

$$z_{obs} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{N}}} = \frac{72.5 - 75}{\frac{16}{\sqrt{120}}} = \frac{-2.5}{1.46} = -1.71$$

  - Calculate the p-value, p_{obs}, using the z-table or EXCEL

    $p = .0436$

  - Make decision by comparing p-value to the α level

    $p_{obs} = .0436 < .05 = p_{crit} = \alpha$

    Reject null hypothesis

- Alternatively, rather than computing an exact p-value, we could have looked up the critical value z_{crit} using the z-table

  $z_{crit} = -1.64$

- Make decision by comparing the observed z-score to the critical z-score

  $z_{obs} = -1.71 < -1.64 = z_{crit}$

  Reject null hypothesis

- Conclude that this sample of high school seniors has significantly lower scores than the population of high school seniors.

- Note 1: Either method of determining significance will result in the exact same decision. With computers, it is easy to calculate exact p-values and they should always be reported.

- Note 2: Now it is up to us, as the researchers, to figure out (and clearly specify) WHY this sample of high school seniors differed from high school seniors in general

- Example 1b: A one-tailed test – take 2
  Suppose that before the data were collected, we had hypothesized that the scores from this sample of high school seniors would be higher than the national average

  - State Null and Alternative Hypotheses
    $$H_0 : \mu \leq 75$$
    $$H_1 : \mu > 75$$

  - Specify α and the sample size
    $$\alpha = .05 \qquad N = 120$$

  - Compute the test statistic z_{obs} based on the sample data
    $$z_{obs} = \frac{72.5 - 75}{\frac{16}{\sqrt{120}}} = \frac{-2.5}{1.46} = -1.71$$

  - Calculate the p-value, p_{obs}, using the z-table or EXCEL
    $$p_{obs} = .9564$$

  - Make decision by comparing p-value to the α level
    $$p_{obs} = .9564 > .05 = p_{crit} = \alpha$$
    Fail to reject / Retain null hypothesis

  - Alternatively, rather than computing an exact *p*-value, we could have looked up the critical value z_{crit} using the z-table
    $$z_{crit} = 1.64$$

  - Make decision by comparing the observed z-score to the critical z-score
    $$z_{obs} = -1.71 < 1.64 = z_{crit}$$

    Fail to reject / Retain null hypothesis

  - Conclude that we do not have enough evidence to claim that this sample of high school seniors is higher than the population of high school seniors.

- Example 1c: A two-tailed test

  Suppose that before the data were collected, we had hypothesized that the scores from this sample of high school seniors differed from the overall population mean (use $\alpha = .05$)?

  - State Null and Alternative Hypotheses

    $$H_0 : \mu = 75$$
    $$H_1 : \mu \neq 75$$

  - Specify α and the sample size

    $\alpha = .05$ $\quad$ $N = 120$

  - Compute the test statistic z_{obs} based on the sample data

    $$z_{obs} = \frac{72.5 - 75}{\frac{16}{\sqrt{120}}} = \frac{-2.5}{1.46} = -1.71$$

  - Calculate the p-value, p_{obs}, using the z-table or EXCEL

    $p(z < z_{obs}) = .0436$

    $p(z > -z_{obs}) = .0436$ $\quad$ $p_{obs} = .0872$

  - Make decision by comparing p-value to the α level

    $p_{obs} = .0872 > .05 = p_{crit} = \alpha$

    Fail to reject / Retain null hypothesis

  - Alternatively, we could calculate the two-tailed critical value z_{crit} using the z-table (For two-tailed test, we need the area beyond the critical value to be equal to .025)

    $z_{crit} = 1.96$

  - Make decision

    $|z_{obs}| = 1.71 < 1.96 = |z_{crit}|$

    Fail to reject / Retain null hypothesis

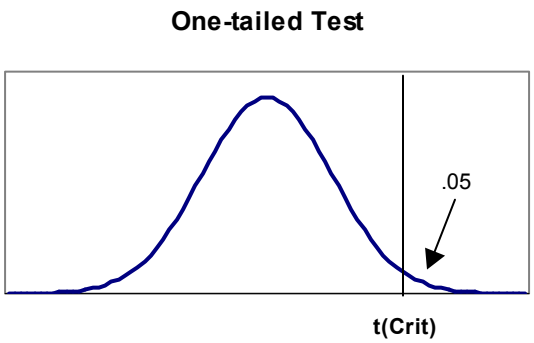
  - Conclude that we do not have enough evidence to claim that this sample of high school seniors differs from the population of high school seniors.

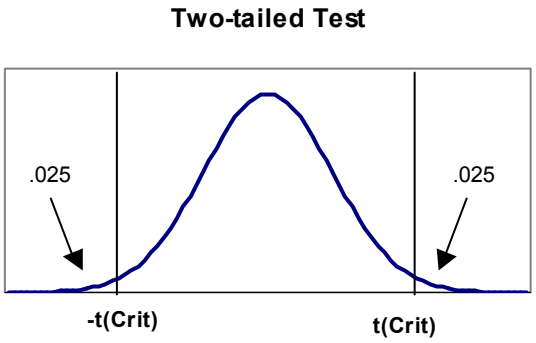
- Summary Table

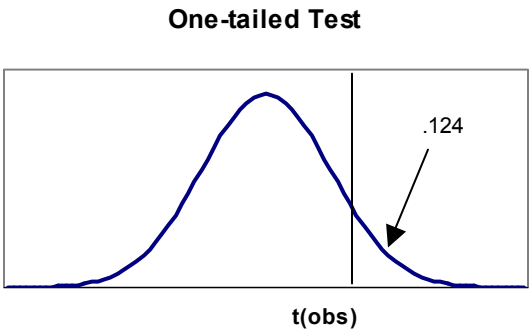
	Two-tailed Test $H_1: \mu \neq 75$	One-tailed Test $H_1: \mu < 75$	One-tailed Test $H_1: \mu > 75$
N	120	120	120
α	.05	.05	.05
z_{obs}	-1.71	-1.71	-1.71
p-value	.0872	.0436	1-.0436
Decision	Fail to Reject	Reject	Fail to Reject

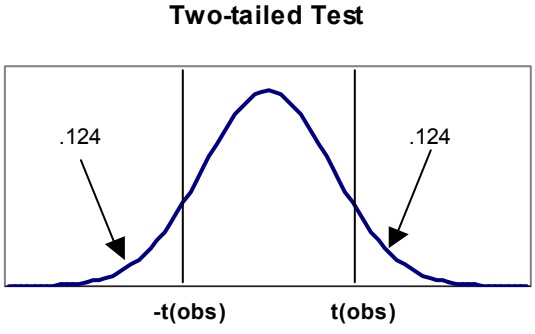
  - This table illustrates why people are skeptical of one-tailed tests!

- A review of the difference between a one tailed and two tailed test:









p = .124

Need to double one-tailed p-value

p = .248

- Example 1d: A two-tailed test with a larger sample
  Suppose that rather than having a sample of size 120, we had a sample of size 160

  - State Null and Alternative Hypotheses
    $H_0 : \mu = 75$
    $H_1 : \mu \neq 75$

  - Specify α and the sample size
    $\alpha = .05 \qquad N = 160$

  - Compute the test statistic z_{obs} based on the sample data

$$z_{obs} = \frac{72.5 - 75}{\frac{16}{\sqrt{160}}} = \frac{-2.5}{1.26} = -1.98$$

  - Calculate the *p*-value, p_{obs}, using the z-table or EXCEL
    $p(z < z_{obs}) = .0239$
    $p(z > -z_{obs}) = .0239 \qquad p_{obs} = .0478$

  - Make decision by comparing p-value to the α level
    $p_{obs} = .0478 < .05 = p_{crit} = \alpha$
    Reject null hypothesis

  - Alternatively, we could calculate the two-tailed critical value z_{crit} using the z-table (For two-tailed test, we need the area beyond the critical value to be .025)
    $z_{crit} = 1.96$

  - Make decision
    $|z_{obs}| = 1.98 > 1.96 = |z_{crit}|$
    Reject null hypothesis

  - Conclude that this sample of high school seniors has significantly different scores than the population of high school seniors.

- Summary Table

	Two-tailed Test $H_1: \mu \neq 75$	Two-tailed Test $H_1: \mu \neq 75$
Raw effect size	-2.5	-2.5
N	120	160
α	.05	.05
z_{obs}	-1.71	-1.98
p-value	.0872	.0478
Decision	Fail to Reject	Reject

- The *p*-value is a function of the difference between means AND the sample size.
- Wouldn't it be nice if we had a measure that did not depend on the sample size?

- Standardized effect sizes:
  - A common standardized measure of effect size: Cohen's *d*

In general, $d = \frac{|M_1 - M_2|}{\sigma_{pooled}}$

For a one-sample z-test, $d = \frac{|\bar{X} - \mu|}{\sigma}$

  - To interpret *d*:

d=0.20	small effect
d=0.50	medium effect
d=0.80	large effect

  - Cohen's *d* is independent of sample size. In other words, increasing the sample size will not (in general) affect *d*.

- Examples 1a-1c:
  - One-sample test with *N = 120*
    (Directionality of test does not matter)

$$d = \frac{|\bar{x} - \mu|}{\sigma} = \frac{|72.5 - 75|}{16} = .16$$

- Example 1d
  - One-sample test with *N = 160*

$$d = \frac{|\bar{x} - \mu|}{\sigma} = \frac{|72.5 - 75|}{16} = .16$$

  - This example would be a "small" effect size
    The sample mean is 0.16 standard deviations below the mean of the population

- An alternative way to interpret *d* scores

Effect Size *d*	Percentage of people in Group1 who would be below the average person in Group 2/the population
0.0	50%
0.1	54%
0.2	58%
0.3	62%
0.4	66%
0.5	69%
0.6	73%
0.7	76%
0.8	79%
0.9	82%
1.0	84%
1.2	88%
1.4	92%
1.6	95%
1.8	96%
2.0	98%
2.5	99%
3.0	99.9%

- An introduction to confidence intervals (CIs)
  - A problem with hypothesis testing is that people tend to only report the point estimates (the estimates of the means). As a reader (and even as the researcher), it can be very easy to forget that there is a distribution of scores with variability
  - Constructing CIs is one way to display the variability in the data
  - In general, a CI is determined by

$$Estimate \pm (Critical\ Value * Standard\ Error\ of\ the\ Estimate)$$

  - How did we arrive at this formula? Let's develop some intuition about this formula
    - For a z-test, we have the following formula for our test statistic:

$$z_{obs} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{N}}}$$

    - Or more generally, the statistic is of the form

$$test_{obs} = \frac{Estimate - Null}{Standard\ error\ of\ estimate}$$

    - At the exact point of rejection, the observed test statistic will equal the critical value

$$test_{crit} = \frac{Estimate - Null}{Standard\ error\ of\ estimate}$$

    - But for a two-tailed test, the critical value can be positive or negative

$$+test_{crit} = \frac{Estimate - Null}{Standard\ error\ of\ estimate} \qquad -test_{crit} = \frac{Estimate - Null}{Standard\ error\ of\ estimate}$$

- Now let's rearrange some terms

$$+test_{crit} = \frac{Estimate - Null}{Standard\ error\ of\ estimate} \qquad -test_{crit} = \frac{Estimate - Null}{Standard\ error\ of\ estimate}$$

$$(Estimate - Null) + (Critical\ Value * Standard\ Error\ of\ the\ Estimate)$$

and

$$(Estimate - Null) - (Critical\ Value * Standard\ Error\ of\ the\ Estimate)$$

- We can combine these two formula to obtain the general formula for a confidence interval

$$(Estimate - Null) \pm (Critical\ Value * Standard\ Error\ of\ the\ Estimate)$$

In many cases, the null value is equal to zero, so the equation is frequently written omitting the null value:

$$Estimate \pm (Critical\ Value * Standard\ Error\ of\ the\ Estimate)$$

- Interpreting the CI can be tricky!
  - The standard interpretation of a $(1-\alpha)\%$ CI is that $(1-\alpha)\%$ of such intervals under repeated sampling contain the population mean
  - The confidence interval is treated as random, changing from sample to sample, and μ is a fixed value

  - There is a connection between the CI and the hypothesis test. The hypothesis test is identical to checking whether the confidence interval includes the value of the null hypothesis.
  - BEWARE! Constructing one-tailed CIs can be tricky!

  - But in a sense, the CI is more informative than the hypothesis test. With a hypothesis test, it is easy to lose a sense of the possible variability of the parameter estimate. With a CI, information about the variability of the estimate is easily accessible.
  - For any hypothesis test, you should always present some display of the variability in the data. A CI is one of the more common ways to do so.

- Example 1a:
  - One-sample, two-tailed hypothesis with $N = 120$
    $z_{crit} = 1.96$

    $$\overline{X}_{obs} \pm \left(z_{crit} * \frac{\sigma}{\sqrt{N}} \right) \Rightarrow 72.5 \pm (1.96 * 1.46) \Rightarrow 72.5 \pm 2.86 \Rightarrow (69.64, 75.36)$$

    OR

    $$\left(\overline{X}_{obs} - \mu_0\right) \pm \left(z_{crit} * \frac{\sigma}{\sqrt{N}} \right) \Rightarrow (72.5 - 75) \pm (1.96 * 1.46) \Rightarrow -2.5 \pm 2.86 \Rightarrow (-5.36, 0.36)$$

  - Interpretation of CIs:

    The first CI provides a CI around the sample mean. This interval includes the null hypothesis $\mu = 75$, which indicates we fail to reject the null hypothesis.

    The second CI provides a CI around the difference between the sample mean and the population mean. This interval includes zero, indicating that zero is a possible value for the difference. Thus, we fail to reject the null hypothesis.

- Example 1d:
  - One-sample, two-tailed hypothesis with $N = 160$
    $z_{crit} = 1.96$

    $$\overline{X}_{obs} \pm \left(z_{crit} * \frac{\sigma}{\sqrt{N}} \right) \Rightarrow 72.5 \pm (1.96 * 1.26) \Rightarrow 72.5 \pm 2.47 \Rightarrow (70.03, 74.97)$$

  - This interval does NOT include the null hypothesis $\mu = 75$, which indicates we reject the null hypothesis

- z-tests: A Final word
  - What if we were not certain that the population had a normal distribution (but we still knew its mean and variance)?
    - The central limit theorem comes to the rescue! Because of the CLT, the sampling distributions of the mean will tend to be normal when the sample size gets large.
    - In these cases, the z-test can be used as an *approximation* to make inferences about the population parameters based on the sample statistics drawn from populations that may or may not be normal.

  - In order to use the z-test, we must know the population variances in advance

  - Unfortunately, we do not always know the variance of the underlying population distribution a priori. In these cases, we will have to estimate these values from the data and rely on different hypothesis tests.

3. One sample t-test

- The z-test is used when the data are known to come from a normal distribution with a known variance. But in practice, we often have no way of knowing the population parameters in advance. Thus we have to rely on estimates of the population parameters, and the t-test.

- The logic of a one-sample t-test
  - Now we need to estimate both the population mean and the population variance
  - We already proved that the sample mean is an unbiased estimator of the population mean

$$\hat{\mu} = \overline{X}$$

  - And we determined the formula for an unbiased estimate of the sample variance:

$$\hat{\sigma}^2 = s^2 = \frac{\sum (x_i - \overline{X})^2}{N-1}$$

  - From our discussion on sampling distributions, we know that estimating the variance is not enough; we need an estimate of variance of the sampling distribution of the mean. When the variance was known, we found:

$$\sigma_{\overline{X}}^2 = \frac{\sigma^2}{N} \qquad \text{or} \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}}$$

    - Now we have an estimator of the standard deviation of a sampling distribution (the standard error). So let's substitute that estimator into the equation:

$$\hat{\sigma}_{\overline{X}} = \frac{\hat{\sigma}}{\sqrt{N}} = \frac{\sqrt{\frac{\sum (x_i - \overline{X})^2}{N-1}}}{\sqrt{N}}$$

- Putting all of this information together, we can construct a test statistic similar to the one sample z-test, but for use in cases where the variance must be estimated from the data

$$t = \frac{\overline{X} - \mu}{\frac{\hat{\sigma}}{\sqrt{N}}}$$

- But this test no longer follows a standard normal distribution, particularly for small samples. W. S. Gossett discovered this problem in the early 1900's. Gossett found a new family of distributions called the t-distribution.
- It can be shown that (under certain conditions):

$$t \sim \frac{\text{estimate of population parameter - expected population parameter}}{\text{estimated standard deviation of the sampling distribution}}$$
$$\text{(usually called the standard error of the distibution)}$$

- Fun facts about the t-distribution:
  - It is leptokurtic
  - As $N \to \infty$, the t-distribution approaches a $N(0,1)$ distribution
  - Here is the density function of the t:

$$f(x_i) = \frac{\Gamma\left[\frac{\nu+1}{2}\right]}{\sqrt{\nu\pi}\left(\frac{\nu}{2}\right)}\left(1+\frac{x_i^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

  - Note that the t distribution depends only on one parameter, ν. This single parameter is called the <u>degrees of freedom</u> of the test. In most of the cases that will interest us, $\nu = n-1$ or $\nu = n-2$

  - In the case of a one-sample t-test, $\nu = n-1$
  - From the point forward, we will refer to ν as *df* (degrees of freedom)

- In order to use the <u>one sample t-test</u>:
  - The population must be normally distributed
  - The population variance is not known in advance
  - The observations are independent and randomly drawn from the population

- Just as for the z-test, we can compute effect sizes and construct confidence intervals.

  - To compute the effect size, we can again use Cohen's *d* or a modification of Cohen's *d*, called Hedges's *g*. For Cohen's *d*, we use the actual population standard deviation; for Hedges's *g*, we use the estimated population standard deviation:

$$g = \frac{|\overline{X} - \mu|}{\hat{\sigma}} \qquad d = \frac{|\overline{X} - \mu|}{\sigma}$$

    - We cannot compute *d* directly, because we do not know σ. But we can compute *d* from *g*:

$$d = g\sqrt{\frac{N}{df}} = g\sqrt{\frac{N}{N-1}}$$

    - What's the difference between *d* and *g*?
      $\Rightarrow$ *g* is descriptive: It describes effect size of the sample
      $\Rightarrow$ *d* is inferential: It describes the effect size of the population

      $\Rightarrow$ However, interpretation of *d* and *g* is the same.

  - For a confidence interval, we can compute a t-critical value based on the alpha level and the appropriate degrees of freedom. Rearranging the formula for the t-test, we obtain a confidence interval around the mean:

$$\overline{X}_{obs} \pm \left(t_{crit} * \frac{\hat{\sigma}}{\sqrt{N}} \right)$$

    or we can obtain a CI around the difference between the sample mean and the (null) hypothesized value

$$\left(\overline{X}_{obs} - \mu\right) \pm \left(t_{crit} * \frac{\hat{\sigma}}{\sqrt{N}} \right)$$

- TV viewing example: In 1995, there were several studies showing that the average American watched 21.22 hours of television per week. A researcher wants to determine if TV viewing has changed. 50 Americans were randomly sampled, and the following data were obtained on average hours spent watching television:

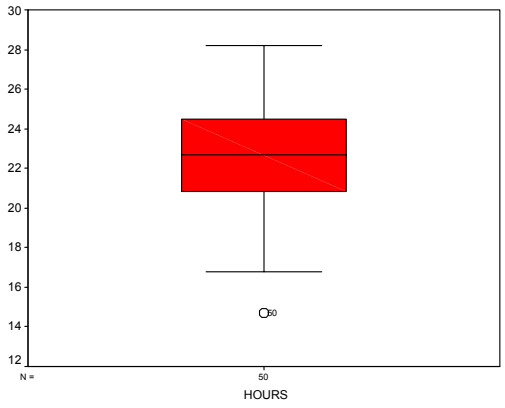
Hours of TV viewing				
21.96	19.38	23.69	26.11	18.82
22.81	21.98	25.79	21.67	24.35
28.18	18.69	21.23	18.37	25.60
23.87	25.11	24.23	20.90	19.51
22.65	20.90	21.20	28.04	16.77
25.39	26.89	21.61	20.14	20.75
23.81	21.74	23.68	23.80	21.40
18.36	24.12	25.40	23.36	26.46
20.20	20.82	21.11	20.76	23.16
22.69	24.51	25.21	24.50	14.68

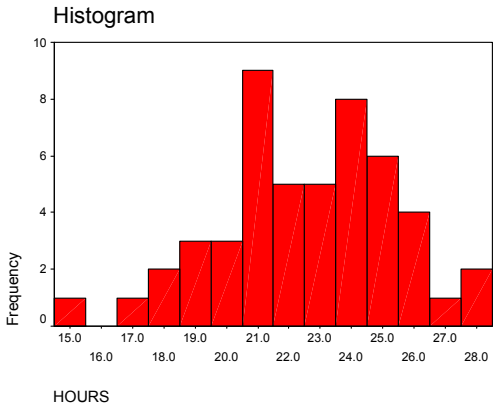
Based on these data, can we claim that television viewing patterns have changed since 1995?

- Step 1: Up to this point, we have always known the distribution of the population in advance. Now, we must look at the data to make sure that our assumptions of the distribution are valid

Always look at the data before jumping into the hypothesis test!
Assumptions we need to check include
  - Are the data normally distributed (or at least symmetrical)?
  - Are there any outliers?

```
EXAMINE VARIABLES=hours
  /PLOT BOXPLOT HISTOGRAM.
```





Descriptives

			Statistic	Std. Error
HOURS	Mean		22.5272	.40286
	95% Confidence Interval for Mean	Lower Bound	21.7176	
		Upper Bound	23.3368	
	5% Trimmed Mean		22.5792	
	Median		22.6700	
	Variance		8.115	
	Std. Deviation		2.84866	
	Minimum		14.68	
	Maximum		28.18	
	Range		13.50	
	Interquartile Range		3.6975	
	Skewness		-.294	.337
	Kurtosis		.103	.662

- It turns out that the t-test is relatively robust to violations of normality. In other words, the t-values and the p-values we obtain from the test are relatively accurate even if the data are not normally distributed
- However, the t-test is NOT robust to violations of symmetry. If the data are not distributed symmetrically around the mean, then the t-values and p-values will be biased

- Thus, we need to check if the data are symmetrical around the mean
  - Is the boxplot symmetrical?
    - Is the median in the center of the box?
    - Do the whiskers extend equally in each direction?
  - Does the histogram look symmetrical?
  - Is the mean approximately equal to the median?
  - Is the coefficient of skewness relatively small?

- We should also be on the look out for outliers – observations that fall far from the main distribution of the data. We would not like our conclusions to be influenced by one point, (or a small number of points). We should not toss out the outliers, but we do need to keep track of them

- In this case, the distribution appears to satisfy all assumptions

- Step 2: Once we have examined the data, and only then, can we conduct the hypothesis test
  - State Null and Alternative Hypotheses

$$H_0 : \mu = 21.22$$
$$H_1 : \mu \neq 21.22$$

  - Specify α and the sample size

$$\alpha = .05 \qquad n = 50$$

  - Compute the test statistic t_{obs} based on the sample data

$$t_{obs} = \frac{22.53 - 21.22}{\frac{2.85}{\sqrt{50}}} = \frac{1.31}{.403} = 3.25$$

  - Calculate the p-value, p_{obs}, based on the appropriate degrees of freedom using the t-table or EXCEL

$$df = n - 1 = 49$$
$$p[t(49) > 3.25] = .0010$$
$$p[t(49) < -3.25] = .0010 \qquad p_{obs} = .0021$$

  - Make decision by comparing p-value to the α level

$$p_{obs} = .0021 < .05 = p_{crit} = \alpha$$

Reject null hypothesis

Americans watch more TV per week now than they did in 1995

  - Alternatively, we could calculate the two-tailed critical value t_{crit} using the t-table. We need to use a two-tailed criteria ($\alpha = .025$ in each tail) with the appropriate degrees of freedom, $df = n - 1 = 49$

$$t_{crit} = 2.01$$

  - Make decision

$$|t_{obs}| = 3.25 > 2.01 = |t_{crit}|$$

Reject null hypothesis

- Calculate an effect size

$$g = \frac{|\overline{X} - \mu|}{\hat{\sigma}} = \frac{|\overline{X} - \mu|}{s} = \frac{|22.53 - 21.22|}{2.85} = .46$$

$$d = .46\sqrt{\frac{50}{49}} = .47$$

- Create confidence intervals

$$\overline{X}_{obs} \pm \left(t_{crit} * \frac{\hat{\sigma}}{\sqrt{N}} \right) \Rightarrow 22.53 \pm \left(2.01 * \frac{2.85}{\sqrt{50}} \right) \Rightarrow 22.53 \pm .810 \Rightarrow (21.72, 23.34)$$

- Using SPSS for a one-sample t-test
  - To use SPSS for a one-sample t-test, we need to have our data entered in a single column.

```
data list free
 /hours.
Begin data.
21.96
22.81
. . .
23.16
14.68
End data.
```

- In the SPSS syntax, we need to specify the DV (hours) and the null hypothesized value, $H_0 : \mu = 21.22$

```
T-TEST  /TESTVAL=21.22
 /VARIABLES=hours.
```

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
HOURS	50	22.5272	2.84866	.40286

One-Sample Test

	Test Value = 21.22					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
HOURS	3.245	49	.002	1.3072	.4976	2.1168

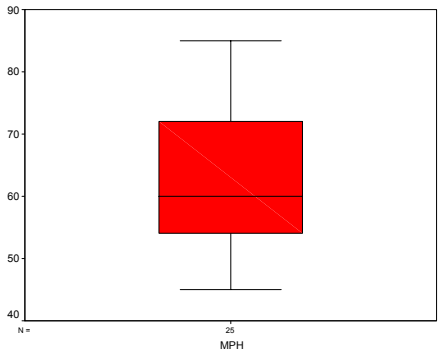
$t(49) = 3.25, p = .002, d = .47$

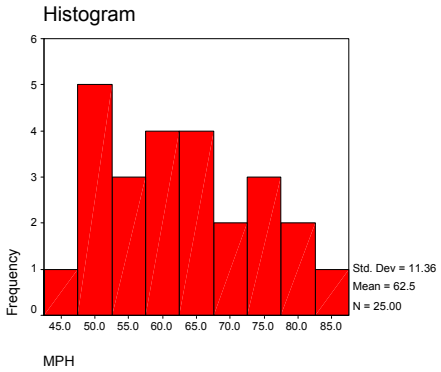
- SPSS computes CIs around the difference between the sample mean and the hypothesized value. If you want a CI around the sample mean, you need to add the (null) hypothesized value to each endpoint of the CI

(21.72, 23.34)

- Example #2: A return to the speeding data. Are people violating the speed limit?

```
EXAMINE VARIABLES=mph
  /PLOT BOXPLOT HISTOGRAM.
```





Descriptives

			Statistic	Std. Error
MPH	Mean		62.5200	2.27165
	95% Confidence Interval for Mean	Lower Bound	57.8315	
		Upper Bound	67.2085	
	5% Trimmed Mean		62.2444	
	Median		60.0000	
	Variance		129.010	
	Std. Deviation		11.35826	
	Minimum		45.00	
	Maximum		85.00	
	Range		40.00	
	Interquartile Range		20.0000	
	Skewness		.380	.464
	Kurtosis		-.859	.902

  - All looks fine to proceed with our hypothesis test

- State Null and Alternative Hypotheses (in this case, a one-tailed hypothesis)

$$H_0 : \mu \leq 55$$
$$H_1 : \mu > 55$$

- Specify α and the sample size

$$\alpha = .05 \qquad n = 25$$

- Compute the test statistic t_{obs} based on the sample data

```
T-TEST /TESTVAL=55
/VARIABLES=mph.
```

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
MPH	25	62.5200	11.35826	2.27165

One-Sample Test

	Test Value = 55					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
MPH	3.310	24	.003	7.5200	2.8315	12.2085

- Calculate the p-value, p_{obs}. SPSS gives us a two-tailed test of significance. We need to adjust the p-value so that it reflects the one-tailed probability
  - We can look up a one-tailed p-value for $t = 3.310$ and $df = 24$ in EXCEL
  - Or we can divide the two-tailed p-value in half
  - Both methods will give the same results

$$t(24) = 3.31, p = .0014$$

- Make decision by comparing p-value to the α level

$$p_{obs} = .00141 < .05 = p_{crit} = \alpha$$

Reject null hypothesis

At this location, people drive faster than the speed limit.

- In general, we do not calculate confidence intervals for one-tailed tests, but we still need to calculate the effect size

$$g = \frac{|\bar{x} - \mu|}{\hat{\sigma}} = \frac{|62.52 - 55|}{11.36} = .66 \qquad d = g\sqrt{\frac{N}{df}} = .66\sqrt{\frac{25}{24}} = .67$$

$t(24) = 3.31, p = .001, d = .67$

4. Two independent samples t-test
   - Now, we would like to compare the means of two independent groups when we do not know the population variance in advance. Let's use something we already know to help us solve this problem:

$$t \sim \frac{\text{estimate of population parameter - expected population parameter}}{\text{estimated standard deviation of the sampling distribution}}$$
$$\text{(usually called the standard error of the distribution)}$$

$$t \sim \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\text{estimated standard deviation of the sampling distribution of } \left(\overline{X}_1 - \overline{X}_2\right)}$$
$$\text{(usually called the standard error of } \overline{X}_1 - \overline{X}_2\text{)}$$

   - We need to describe the sampling distribution of $\overline{X}_1 - \overline{X}_2$
     - From probability theory we know that the variance of the difference of two variables is:

$$Var(\overline{X}_1 - \overline{X}_2) = Var(\overline{X}_1) + Var(\overline{X}_2) - 2\,\mathrm{cov}(\overline{X}_1, \overline{X}_2)$$

     - To make our life much simpler, let's make two assumptions
       1. Assume $\overline{X}_1$ and $\overline{X}_2$ are <u>independent</u>

$$\begin{aligned} Var(\overline{X}_1 - \overline{X}_2) &= Var(\overline{X}_1) + Var(\overline{X}_2) + 0 \\ &= Var\left(\frac{X_{11} + \ldots + X_{1n_1}}{n_1}\right) + Var\left(\frac{X_{21} + \ldots + X_{2n_2}}{n_2}\right) \\ &= \frac{1}{n_1^2}\left[Var(X_{11}) + \ldots + Var(X_{1n_1})\right] + \frac{1}{n_2^2}\left[Var(X_{21}) + \ldots + Var(X_{2n_2})\right] \\ &= \frac{1}{n_1^2}\left[n_1 Var(X_1)\right] + \frac{1}{n_2^2}\left[n_2 Var(X_2)\right] \\ &= \frac{Var(X_1)}{n_1} + \frac{Var(X_2)}{n_2} \end{aligned}$$

2. Assume $Var(X_1) = Var(X_2)$ (Homogeneity of variances)

$$Var(\overline{X}_1 - \overline{X}_2) = \frac{Var(X)}{n_1} + \frac{Var(X)}{n_2}$$
$$= Var(X)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

- This equation does not look so bad. But we will have to remember these two assumptions we made!
- To compute the estimate of $Var(X)$, we can combine (or pool) the individual estimates of $Var(X_1)$ and $Var(X_2)$:

$$Var(X) = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

$$s_{pooled} = \sqrt{\frac{SS_1 + SS_2}{(n_1 + n_2 - 2)}}$$

- Now we have a test statistic we can work with:

$$t_{obs} = \frac{estimate}{std\ error\ of\ estimate}$$
$$= \frac{\overline{X}_1 - \overline{X}_2}{s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$df = n_1 + n_2 - 2$$

- The reason for such great detail is to highlight the importance of the assumptions for the two-sample t-test:
  - Observations are normally distributed in both population 1 and population 2 (or at least symmetrically distributed)
  - The variances of the two populations are unknown, but equal (Equality of variances)
  - The observations are independent and randomly drawn from the population
  - The sample of observations from population 1 is independent from the sample of observations from population 2

- Effect sizes for a two-independent sample t-test
  - Just like for the one-sample test, we have a choice between Cohen's *d* and Hedges's *g*:

$$g = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{s_{pooled}} \qquad d = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sigma}$$

$$g = \frac{2t}{\sqrt{N}} \qquad d = \frac{2t}{\sqrt{df}}$$

  - For a two-sample test, we can also use *r* as a measure of effect size:

$$r = \frac{t}{\sqrt{t^2 + df}} \qquad r = \frac{d}{\sqrt{d^2 + 4}}$$

    ⇒ *r* is interpreted a correlation coefficient – the correlation between the IV and the DV

    ⇒ Rules of thumb for interpreting *r*

r = .1	small effect
r = .3	medium effect
r = .5	large effect

    ⇒ In general, I find the interpretation of *d* or *g* to be more straightforward than *r*. All are commonly accepted; you should use the measure you understand and can interpret best.

  - All effect size formulae perform optimally with equal n and homogeneous variances. Adjustments are available for these other situations (see Rosethal, Rosnow, & Rubin, 2000)

- You may also wish to construct confidence intervals
  - Recall the confidence interval is:

$$Estimate \pm (Critical\ Value * Standard\ Error\ of\ the\ Estimate)$$

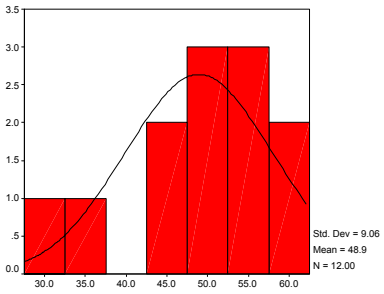
$$\overline{X}_1 - \overline{X}_2 \pm \left(t_{crit} * s_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

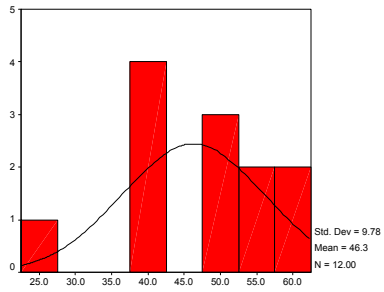
- Example #1: A sleep deprivation example: To investigate the effect of sleep deprivation on memory, a psychologist studies 24 individuals. 12 participants in the normal sleep group receive a normal amount of sleep (8 hours) before going to the lab, while the 12 participants in the sleep deprivation group are only allowed 5 hours of sleep. All participants are then given a recall task.

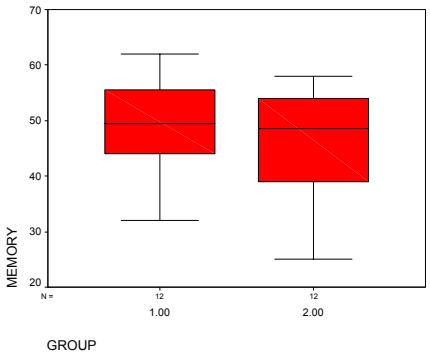
Control		Sleep Deprived	
55	58	48	55
43	45	38	40
51	48	53	49
62	54	58	50
35	56	36	58
48	32	42	25

- Step 1: Look at data and check assumptions. We need to check if:
  - Each group is normally/symmetrically distributed
  - If the variances of the two groups are equal

```
EXAMINE  VARIABLES=memory BY group
  /PLOT BOXPLOT HISTOGRAM.
```







Descriptives

	GROUP			Statistic	Std. Error
MEMORY	1.00	Mean		48.9167	2.61539
		95% Confidence Interval for Mean	Lower Bound	43.1602	
			Upper Bound	54.6731	
		5% Trimmed Mean		49.1296	
		Median		49.5000	
		Variance		82.083	
		Std. Deviation		9.05999	
		Minimum		32.00	
		Maximum		62.00	
		Range		30.00	
		Interquartile Range		12.2500	
		Skewness		-.601	.637
		Kurtosis		-.255	1.232

Descriptives

	GROUP			Statistic	Std. Error
MEMORY	2.00	Mean		46.0000	2.89200
		95% Confidence Interval for Mean	Lower Bound	39.6348	
			Upper Bound	52.3652	
		5% Trimmed Mean		46.5000	
		Median		48.5000	
		Variance		100.364	
		Std. Deviation		10.01817	
		Minimum		25.00	
		Maximum		58.00	
		Range		33.00	
		Interquartile Range		16.0000	
		Skewness		-.697	.637
		Kurtosis		.048	1.232

- We will develop some more refined checks of these assumptions in the near future, but for now, all appears satisfactory.
  (There may be some concern about the slight negative skew of both distributions, but it is not sufficient enough to be troublesome.)

- Now, we can conduct our significance test:
  - State Null and Alternative Hypotheses

$$H_0 : \mu_C = \mu_{SD} \quad \text{or} \quad H_0 : \mu_C - \mu_{SD} = 0$$
$$H_1 : \mu_C \neq \mu_{SD} \qquad H_1 : \mu_C - \mu_{SD} \neq 0$$

  - Specify α and the sample size

$$\alpha = .05 \qquad n_C = 12 \qquad n_{SD} = 12$$

  - Compute the test statistic based on the sample data

$$t_{obs} = \frac{\left(\overline{X}_C - \overline{X}_{SD}\right) - 0}{s_{pooled}\sqrt{\frac{1}{n_C} + \frac{1}{n_{SD}}}}$$

    - In SPSS, we need to specify the IV (group) and the DV (memory)

```
T-TEST GROUPS=group
 /VARIABLES=memory.
```

Group Statistics

	GROUP	N	Mean	Std. Deviation	Std. Error Mean
MEMORY	1.00	12	48.9167	9.05999	2.61539
	2.00	12	46.0000	10.01817	2.89200

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
MEMORY	Equal variances assumed	.256	.618	.748	22	.462	2.9167	3.89922	-5.16982	11.00315
	Equal variances not assumed			.748	21.781	.462	2.9167	3.89922	-5.17453	11.00786

    - We determined that the variances of the two groups were equal, so we read the "equal variances assumed" line:

$$t(22) = 0.75, p = .46$$

- To compute the t-statistic by hand, we need to calculate the pooled estimate of the standard deviation

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}} = \sqrt{\frac{(11)9.06^2 + (11)10.02^2}{22}} = 9.5509$$

$$t_{obs} = \frac{\left(\overline{X}_C - \overline{X}_{SD}\right)}{s_{pooled}\sqrt{\frac{1}{n_C} + \frac{1}{n_{SD}}}} = \frac{(48.92 - 46.00)}{9.5509\sqrt{\frac{1}{12} + \frac{1}{12}}} = \frac{2.92}{3.899} = 0.748$$

- If the null hypothesized value is not zero, then you will need to adjust the observed t-statistic by hand (SPSS assumes the null hypothesis is that the two means are equal). Alternatively, you can use EXCEL, which allows you to enter a null value.

t-Test: Two-Sample Assuming Equal Variances

	Variable 1	*Variable 2*
Mean	48.91667	46
Variance	82.08333	100.3636
Observations	12	12
Pooled Variance	91.22348	
Hypothesized Mean Difference	0	
df	22	
t Stat	0.748013	
P(T<=t) one-tail	0.231187	
t Critical one-tail	1.717144	
P(T<=t) two-tail	0.462374	
t Critical two-tail	2.073875	

Using EXCEL also gives you the pooled variance, s^2_{pooled}, which we can use to compute the effect size *g*.

- Make decision

$p_{obs} = .462 > .05 = p_{crit} = \alpha$

$|t_{obs}(22)| = .748 < 2.07 = t_{crit}(22)$

Fail to reject null hypothesis
Conclusion?

- Create confidence intervals of the difference

$$\overline{X}_1 - \overline{X}_2 \pm \left(t_{crit} * s_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

$$2.92 \pm (2.074 * 3.899) \Rightarrow (-5.17, 11.00)$$

  - This CI matches the CI from SPSS

- Calculate effect size

$$g = \frac{\overline{X}_1 - \overline{X}_2}{s_{pooled}} = \frac{2.92}{9.55} = .306 \qquad r = \frac{.748}{\sqrt{.748^2 + 22}} = .157$$

$$g = \frac{2t}{\sqrt{N}} = \frac{2 * .748}{\sqrt{24}} = .306$$

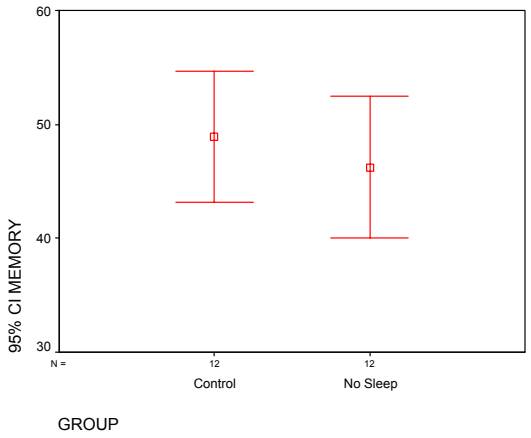
- Now that we have data (with unknown population parameters), we should ALWAYS present a graphical display of the data and that graph should ALWAYS display the variability in the data

- You have two (or three) choices of how to display the variability in the data
  - The mean displayed in the center of the confidence interval
  - The mean $\pm$ 2 standard errors of the mean

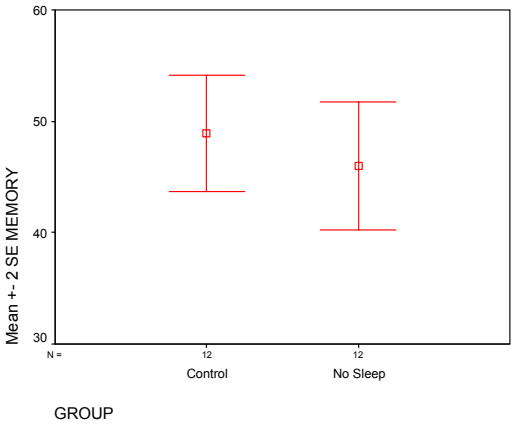
  - Some people (but not SPSS) prefer
    The mean $\pm$ 1 standard error of the mean

- In SPSS

GRAPH /ERRORBAR(CI 95)
=memory by group.



GRAPH/ERRORBAR (STERROR 2)
=memory by group.



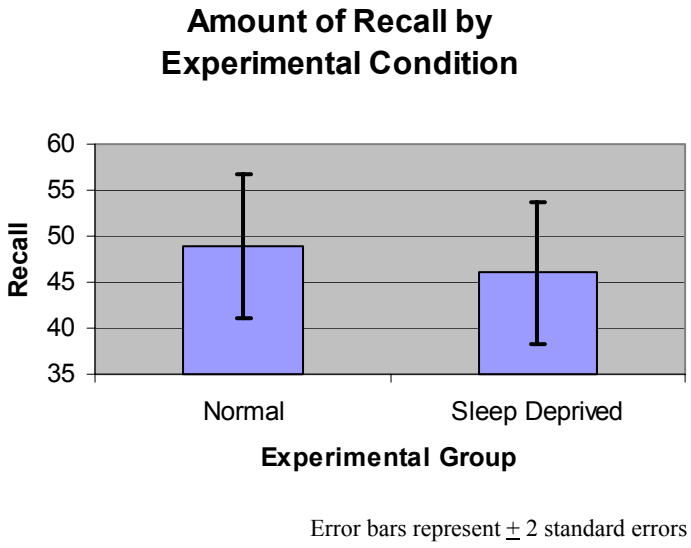
- Should you graph two standard error bars or half-widths of confidence intervals?

$$\text{Standard error of the mean} = 2 * s_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\text{Confidence interval half-width} = t_{crit} * s_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- As the degrees of freedom get larger, for a two-tailed test with $\alpha = .05$, $t_{crit} \Rightarrow 1.96$.
- Thus for large samples, the two produce nearly identical results.
- Use of standard errors is most common. Regardless of your preference, be sure to label your error bars!

- SPSS's graphs tend to be ugly and of non-journal quality. You should be familiar with a graphing program to produce journal quality graphs (for the most part, EXCEL is sufficient).



5. Comparing two group means when assumptions are violated

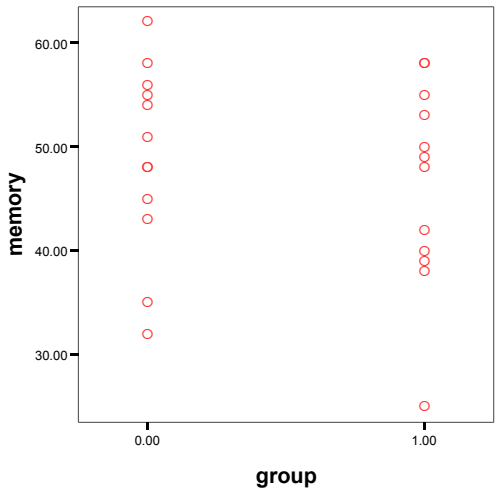
- The Two-sample t-test when things go wrong!
  Recall the assumptions of the two sample t-test:
  - Observations are normally distributed (or at least symmetrically distributed)
  - Equality of variances
  - Independence of observations and populations

- What we need to look for in our data are:
  - Normally distributed data (or symmetric data)
  - Equality of variances
  - Outliers

- Exploratory Data Analysis (EDA) techniques to test assumptions – statistical tests can wait:
  - To check for normality/symmetry
    - Examine mean and median
    - Coefficients of skewness and kurtosis
    - Histograms (should be performed for each group!)
    - Boxplots
  - To check for equality of variances
    - Boxplots
    - Scatter plots
  - To check for outliers
    - Boxplots
    - Histograms
    - Scatter plots

  - We have already examined all of these except for the scatterplot

```
GRAPH
  /SCATTERPLOT(BIVAR)=group WITH memory .
```



- Example #1: A simulation of the effect of outliers' asymmetry on estimates of the mean:

Let $X_1 \sim N(2,1)$ and Let $X_2 \sim N(6,9)$

Let Y = X_1 with probability $1-\varepsilon$
X_2 with probability ε

Let ε = .00 (no outliers)
.01 (2 outliers per 200)
.025 (5 outliers per 200)
.05 (10 outliers per 200)
.10 (20 outliers per 200)

Let n=200 and Let α = .05
Calculate $\overline{Y}$ and a CI for $\overline{Y}$

ε	$\hat{\mu}$	Confidence Coverage (Should be 95%)
0.00	2.00042	95.3
0.01	2.04088	91.5
0.025	2.09849	82.7
0.05	2.20703	56.3
0.10	2.29815	18.5

How would this affect a more realistic sample size of 30?

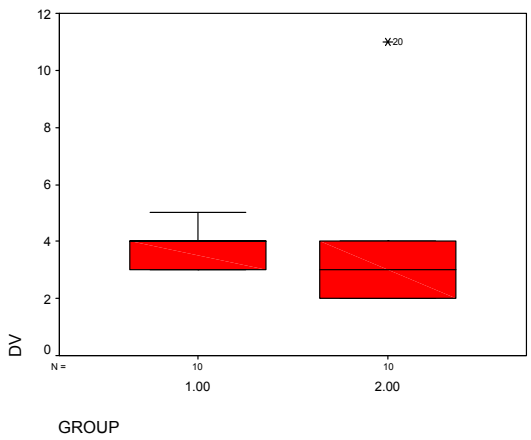
For ε = .00 (no outliers) → no outliers
.01 (2 outliers per 200) → 0.3 outliers per 30
.025 (5 outliers per 200) → 0.75 outliers per 30
.05 (10 outliers per 200) → 1.5 outliers per 30
.10 (20 outliers per 200) → 3.0 outliers per 30

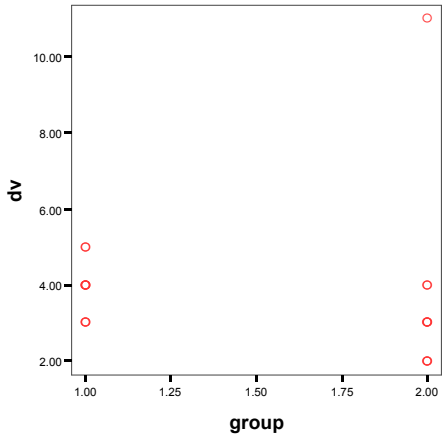
- Example #2: A data-based example of the effect of an outlier:

Group 1	Group 2
3	2
4	3
5	4
4	3
3	2
4	3
5	4
4	3
3	2
4	11

  - Group 1 is always 1 unit larger than Group 2 except for one observation
  - First, let's look at the data:

```
EXAMINE  VARIABLES=DV  BY group
  /PLOT BOXPLOT.
```





Descriptives

	GROUP			Statistic	Std. Error
DV	1.00	Mean		3.9000	.23333
		95% Confidence Interval for Mean	Lower Bound	3.3722	
			Upper Bound	4.4278	
		5% Trimmed Mean		3.8889	
		Median		4.0000	
		Variance		.544	
		Std. Deviation		.73786	
		Minimum		3.00	
		Maximum		5.00	
		Range		2.00	
		Interquartile Range		1.2500	
		Skewness		.166	.687
		Kurtosis		-.734	1.334
	2.00	Mean		3.7000	.84393
		95% Confidence Interval for Mean	Lower Bound	1.7909	
			Upper Bound	5.6091	
		5% Trimmed Mean		3.3889	
		Median		3.0000	
		Variance		7.122	
		Std. Deviation		2.66875	
		Minimum		2.00	
		Maximum		11.00	
		Range		9.00	
		Interquartile Range		2.0000	
		Skewness		2.725	.687
		Kurtosis		7.991	1.334

- Not only do we have a problem with an outlier, but as is often the case, outliers lead to other problems as well
  - The variances of the two groups are very different

    $\hat{\sigma}_1^2 = 0.544$ $\hat{\sigma}_2^2 = 7.122$
  - Group 2 has a strong positive skew (is non-symmetrical)

- When good data go bad, what can we do?
  - Check the data
  - Ignore the problem
  - Transform the variable
  - Perform a test that does not require the assumption
  - Use a non-parametric test
  - Use robust estimators of the mean and variance

- Option 1: Check the data
  - Make sure that the outlier is a true data point and not an error

- Option 2: Ignore the outlier/heterogeneity

```
T-TEST GROUPS=group
 /VARIABLES=DV.
```

Group Statistics

	GROUP	N	Mean	Std. Deviation	Std. Error Mean
DV	1.00	10	3.9000	.73786	.23333
	2.00	10	3.7000	2.66875	.84393

Independent Samples Test

		t-test for Equality of Means						
							95% Confidence Interval of the Difference	
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
DV	Equal variances assumed	.228	18	.822	.2000	.87560	-1.63956	2.03956

$$t(18) = .23, p = .82$$

This is not a very wise choice!

- Option 3: Transform the variable
  - We will cover the details of transformation in the context of ANOVA

- Option 4: Perform a test that does not require the assumption: Welch's separate variance two sample t-test
  - Similar to the two-sample t-test, but does not make the "simplifying" homogeneity of variance assumption
  - Computation is similar to t-test with 2 exceptions
    - No pooling to estimate the variance
    - Degrees of freedom are "adjusted" to take into account the unequal variances

$$t_{obs} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)c^2 + (n_1 - 1)(1 - c^2)}$$

$$c = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

  - Notes:
    - If $n_1 = n_2$, then Welch's t will equal the uncorrected t (but the degrees of freedom of the tests may be different)
    - If $n_1 = n_2$ and if $s_1 = s_2$, then Welch's t will give the exact same result as the uncorrected t.

    - Welch's t provides an unbiased estimate, but it is less efficient than the uncorrected t-test (and this has slightly less power). However, if the population variances are unequal, the uncorrected t-will give a biased result.
    - Given that we can never be sure that variances are equal, it would be reasonable to recommend always using Welch's t-test. The slight decrease in power when the variances are equal will be offset by the fact that the test will always give unbiased results and will maintain the true alpha rate near .05.

- Luckily SPSS will do the dirty work for you!

```
T-TEST GROUPS=group
 /VARIABLES=DV.
```

Independent Samples Test

		t-test for Equality of Means						
							95% Confidence Interval of the Difference	
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
DV	Equal variances assumed	.228	18	.822	.2000	.87560	-1.63956	2.03956
	Equal variances not assumed	.228	10.368	.824	.2000	.87560	-1.74160	2.14160

$t(10.37) = .228, p = .824$

In this case, we have more problems than unequal variances. We still require normality/symmetry for this test. So in this case, the Welch's t-test is not a very good option either!

- Option 5: Perform a non-parametric test

  - In general, non-parametric tests:
    - Make no assumptions about the distribution of the data
    - Reduce the effect of outliers and heterogeneity of variance
    - Are not as powerful as parametric alternatives when the assumptions of the parametric tests are satisfied
    - Can be used for ordinal data

  - By definition, non-parametric tests to not estimate population parameters
    - There are no estimates of variance/variability
    - There are no confidence intervals
    - There are generally fewer measures of effect size available

  - Mann-Whitney U test (Wilcoxon Rank-Sum test)
    - Equivalent to performing the independent samples t-test on the ranks of the data (instead of the raw data)
    - Not as powerful as the t-test (because it ignores the interval nature of the data)

```
RANK VARIABLES=dv.

      From      New
      variable  variable  Label
      --------  --------  -----

      DV        RDV       RANK of DV
```

```
T-TEST GROUPS=group
  /VARIABLES=rdv.
```

Group Statistics

	GROUP	N	Mean	Std. Deviation	Std. Error Mean
RANK of DV	1.00	10	12.80000	4.385582	1.386843
	2.00	10	8.20000	6.033241	1.907878

Independent Samples Test

		t-test for Equality of Means						
							95% Confidence Interval of the Difference	
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
RANK of DV	Equal variances assumed	1.950	18	.067	4.60000	2.358672	-.355386	9.555386

- On the ranked data, we find a marginally significant effect such that the scores of group 1 are larger than the scores of group 2, $t(18) = 1.95$, $p = .067$.

- The Mann-Whitney U test can be performed directly in SPSS

```
NPAR TESTS M-W = dv by group (1,2).
```

Ranks

	GROUP	N	Mean Rank	Sum of Ranks
DV	1.00	10	12.80	128.00
	2.00	10	8.20	82.00
	Total	20		

Test Statistics[b]

	DV
Mann-Whitney U	27.000
Wilcoxon W	82.000
Z	-1.821
Asymp. Sig. (2-tailed)	.069
Exact Sig. [2*(1-tailed Sig.)]	.089[a]

a. Not corrected for ties.

b. Grouping Variable: GROUP

$$U = 27.00, p = .089$$

- These two methods give nearly identical results. However, you should report the results of the U test and not the t-test on the ranks.
- The outlier does not influence this test very much. This test would be a reasonable option to analyze this data without tossing the outlier.

- Option 6: (For exploratory purposes only) Use estimates of central tendency and variability that are robust

Central tendency	Median M-estimators
Variability	IQR Median Absolute Deviation (MAD)

```
EXAMINE  VARIABLES=DV  BY group
  /MESTIMATORS  HUBER(1.339) TUKEY(4.685).
```

M-Estimators

	GROUP	Huber's M-Estimator[a]	Tukey's Biweight[b]
DV	1.00	3.8546	3.8675
	2.00	3.0377	2.8810

a. The weighting constant is 1.339.

b. The weighting constant is 4.685.

Descriptives

	GROUP		Statistic
DV	1.00	Mean	3.9000
		Median	4.0000
		Interquartile Range	1.2500
	2.00	Mean	3.7000
		Median	3.0000
		Interquartile Range	2.0000

We need to compute MAD by hand (or in EXCEL)

$MAD_1 = Med\left\{\left|x_{1i} - Med_{x_1}\right|\right\}$ $\qquad$ $MAD_2 = Med\left\{\left|x_{2i} - Med_{x_2}\right|\right\}$

	Group 1	Group2	Pooled
Median	4.0	3.0	
IQR	1.25	2.0	1.625
M(Tukey)	3.8675	2.8810	
M(Huber)	3.8546	3.0377	
MAD	0.5	1	0.75

To pool the variability estimates, we can adapt the formula for s_{pooled}

$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

$$MAD_{pooled} = \frac{(n_1 - 1)MAD_1 + (n_2 - 1)MAD_2}{(n_1 - 1) + (n_2 - 1)} \qquad IQR_{pooled} = \frac{(n_1 - 1)IQR_1 + (n_2 - 1)IQR_2}{(n_1 - 1) + (n_2 - 1)}$$

$$t_{obs} = \frac{estimate}{std\ error\ of\ estimate}$$
$$= \frac{\overline{X}_1 - \overline{X}_2}{s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- Let's construct some tests using our robust estimators
- Test 1: Use median and IQR

$$K1_{obs}(18) = \frac{Med_1 - Med_2}{.74 * IQR_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{1.0}{.5367} = 1.86$$

- Test 2: Use Tukey Biweight and MAD

$$K2_{obs}(18) = \frac{TukeyM_1 - TukeyM_2}{MAD_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{0.9865}{.3354} = 2.941$$

- Test 3: Use Huber estimator and MAD

$$K3_{obs}(18) = \frac{HuberM_1 - HuberM_2}{MAD_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{0.8298}{.3354} = 2.474$$

- If we assume that each of these tests follow an approximate t-distribution, we can compute *p*-values.

  $K1(18) = 1.85, p = .08$

  $K2(18) = 2.94. p = .009$

  $K3(18) = 2.47, p = .024$

- "It is better to have an approximate solution to the right answer than the correct solution to the wrong answer." – John Tukey
- These tests should be used for exploratory purposes only. The details of these tests have not been completely worked out. However, robust methods are gaining a foothold and it is likely that robust tests similar to these will be popular in the future.

o Importantly, in none of these solutions did we toss the "bad" data point.

6. Pearson Chi-square test of independence

- A non-parametric test used when dependent variable is categorical (measured on a nominal scale)
- It is a test of the independence of the levels of 1 or more factors when all the factors are nominal.
  - We calculate the frequencies we would have observed in each cell if the factors and factor levels were independent – the frequencies expected under the assumption of independence.
  - We compare the observed frequencies to the expected frequencies:

$$\chi^2 = \sum_{Cells} \frac{\left(f_{observed} - f_{expected}\right)^2}{f_{expected}}$$

- The Chi-square test follows a Chi-square distribution.
- The only assumption is that data be independently sampled from the population.

- A one factor Chi-square test of independence
  - A one factor Chi-square test is relatively rare
  - Tests if the observed frequencies are equally distributed over the levels of the factor

- Example #1: It has been suggested that admission to psychiatric hospitals may vary by season. One hospital admitted 100 patients last year with the following distribution:

	Season			
	Spring	Summer	Fall	Winter
Observed	30	40	20	10

Do hospital admissions vary by season?

- State Null and Alternative Hypotheses

H_0 : Hospital admission is independent of season

H_1 : Hospital admission is NOT independent of season

- Compute the test statistic based on the sample data
  - We have the observed frequencies
  - We need to compute the expected frequencies under the null hypothesis. (That is, the frequencies we would have observed if hospital admissions were independent of season)

  If hospital admissions were independent of season, then hospital admissions would be distributed equally over the 4 seasons

$$f_e = \frac{N}{a} = \frac{100}{4} = 25$$

Where N is the total number of observations
a is the number of levels of the factor

	Season			
	Spring	Summer	Fall	Winter
Expected	**25**	**25**	**25**	**25**
Observed	30	40	20	10

- Now we can compute the test statistic

$$\chi^2 = \sum_{Cells} \frac{\left(f_{observed} - f_{expected}\right)^2}{f_{expected}}$$

$$df = a - 1$$

$$\chi^2 = \sum_{Cells} \frac{\left(f_{observed} - f_{expected}\right)^2}{f_{expected}} = \frac{(30-25)^2}{25} + \frac{(40-25)^2}{25} + \frac{(20-25)^2}{25} + \frac{(10-25)^2}{25} = 20$$

$$df = a - 1 = 3$$

- To determine significance, we can:

  Look up the significance level using EXCEL

  $p_{obs} = .00017 < .05 = p_{crit} = \alpha$

  $\chi^2(3) = 20, p < .001$

  Or look up the critical value using a Chi-square table with $\alpha = .05$

  $\chi^2_{obs}(3) = 20 > 7.81 = \chi^2_{crit}(3)$

  $\chi^2(3) = 20, p < .05$

  We reject the null hypothesis and conclude that hospital admission varies by season.

  Note that our test is not focused enough to permit a more specific conclusion. For example, we cannot state that admissions are greater in the fall than in the winter.

- A two factor Chi-square test of independence
  - A two factor Chi-square test is the most common application of the Chi-square test.
  - Tests if the observed frequencies are independent across the levels of the factor

- Example #2: Belief in the Afterlife
  - A researcher wondered if belief in an afterlife differed by gender. She obtained a random sample of 1091 individuals and with the following data:

Observed Counts

	Belief in Afterlife	
Gender	Yes	No
Females	435	147
Males	375	134

Does belief in an afterlife vary by gender?

- State Null and Alternative Hypotheses

  H_0 : Belief in an afterlife is independent of gender

  H_1 : Belief in an afterlife is NOT independent of gender

- Now, we need to compute the observed frequencies under the null hypothesis (the assumption of independence).

  - Step 1: Calculate the Row and Column totals

	Belief in Afterlife		
Gender	Yes	No	
Females	435	147	582
Males	375	134	509
	810	281	1091

  - Calculate the Expected Cell Frequencies if the data were independent:

	Belief in Afterlife		
Gender	Yes	No	
Females			582
Males			509
	810	281	1091

$$f_e = \frac{RowTotal * ColumnTotal}{N}$$

  - Expected frequencies:

	Belief in Afterlife		
Gender	Yes	No	
Females	432	150	582
Males	378	131	509
	810	281	1091

- Now calculate Pearson Chi-square statistic

$$\chi^2 = \sum_{Cells} \frac{\left(f_{observed} - f_{expected}\right)^2}{f_{expected}}$$

$$df = (a-1)(b-1)$$

a = Number of levels of the first factor
b = Number of levels of the second factor

- $\frac{Observed}{Expected}$

Gender	Belief in Afterlife: Yes	Belief in Afterlife: No
Females	$435/432$	$147/150$
Males	$375/378$	$134/131$

$$\chi^2 = \sum_{Cells} \frac{\left(f_{observed} - f_{expected}\right)^2}{f_{expected}} = \frac{9}{432} + \frac{9}{150} + \frac{9}{378} + \frac{9}{131} = .173$$
$$= 0.173$$

$$df = (\#\ of\ rows - 1)(\#\ of\ columns - 1)$$
$$= 1$$

- To determine significance, $\alpha = .05$, we can:

  Look up the significance level using EXCEL

  $\chi^2(1) = 0.17, p = .68$

  $p_{obs} = .677 > .05 = p_{crit} = \alpha$

  Or look up the critical value using a Chi-square table with $\alpha = .05$

  $\chi^2{}_{obs}(1) = 0.173 < 3.84 = \chi^2{}_{crit}(1)$

  $\chi^2(1) = 0.17, p > .05$

  We retain the null hypothesis. There is not sufficient evidence to conclude that belief in an afterlife varies by gender.

WARNING: Be careful of round-off error when calculating Chi-square tests by hand!

- Example #3: Favorite sport
  Imagine that you read an article claiming that 300 men and 250 women were interviewed. Of the men, 55% said that baseball was their favorite sport, while 30% of the women said that baseball was their favorite sport. You think – hmmm, I wonder if this is a significant difference.

  H_0 = The proportion of people who state baseball is their favorite sport is equal among men and women
  H_1 = The proportion of people who state baseball is their favorite sport is NOT equal among men and women

  Or

  H_0 = Rating baseball as one's favorite sport is independent of one's gender
  H_1 = Rating baseball as one's favorite sport is NOT independent of one's gender

  - Step 1: Convert the data into count form for a contingency table

    .55*300 = 165 men said baseball was their favorite sport
    300-165 = 135 men said baseball was NOT their favorite sport

    .30*250 = 75 women said baseball was their favorite sport
    250-75 = 175 women said baseball was NOT their favorite sport

Observed

	Baseball Favorite?	
Gender	Yes	No
Females	75	175
Males	165	135

- Step 2: Determine the expected frequencies under the null hypothesis, and then calculate the test-statistic. Let's skip the hand calculations and proceed to use SPSS.

  Yucky method
  - We have n=550, so we need to enter 550 rows of data

  Simple method
  - Enter one row for each cell and the count in that cell

```
0 0 175
0 1 75
1 0 135
1 1 155
```

  Column 1: 0 = Female
  1 = Male
  Column 2: 0 = No, baseball is not my favorite sport
  1 = Yes, baseball is my favorite sport
  Column 3: Count associated with the specific cell

```
DATA LIST FREE
  /gender baseball count.
BEGIN DATA.
0 0 175
0 1 75
1 0 135
1 1 165
END DATA.

VALUE LABELS
   gender  0 'Female' 1 'Male'
   /baseball  0 'No' 1 'Yes' .
EXECUTE.
```

  - The trick is to weight the analysis by the cell counts

```
WEIGHT BY count .
CROSSTABS
  /TABLES=gender  BY baseball
  /STATISTIC=CHISQ.
```

GENDER * BASEBALL Crosstabulation

Count

		BASEBALL		
		No	Yes	Total
GENDER	Female	175	75	250
	Male	135	165	300
Total		310	240	550

Check this table to make sure you have entered the data correctly!

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	34.652[b]	1	.000		
Continuity Correction[a]	33.643	1	.000		
Likelihood Ratio	35.213	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	34.589	1	.000		
N of Valid Cases	550				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 109.09.

- Read line labeled 'Pearson' or 'Pearson Chi-square'

  $\chi^2(1) = 34.65, p < .01$

- Reject null hypothesis and conclude that the proportion of people who rate baseball as their favorite varied by gender.

- Compute the effect size of the two-way chi-square test, phi

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

A small effect $\phi = .10$
A medium effect $\phi = .30$
A large effect $\phi = .50$

$$\phi = \sqrt{\frac{34.652}{550}} = .25$$

$\chi^2(1) = 34.65, p < .01, \phi = .25$

- What can go wrong in a chi-square test:
  - The χ^2 statistic only has a chi-square distribution if the expected cell sizes are "large" where cell sizes > 5 are considered large.
  - For a 2*2 table, the solution for small cell sizes is to use Fisher's exact test (It is based on exact probabilities calculated from the hypergeometric distribution)

- Extending the Chi-square beyond a 2*2:
  - Consider a 2*3 example: Gender and Party Affiliation

	Party Affiliation		
Gender	Democrat	Independent	Republican
Female	279	73	225
Male	165	47	191

H_0 : Party affiliation is independent of gender

H_1 : Party affiliation varies by gender

  - Calculate expected frequencies, based on the null hypothesis

	Party Affiliation			
Gender	Democrat	Independent	Republican	
Female	279	73	225	577
	(261.4)	(70.7)	(244.9)	
Male	165	47	191	403
	(182.6)	(49.3)	(171.1)	
	444	120	416	980

$$f_e = \frac{RowTotal * ColumnTotal}{N}$$

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected} = 7.010$$

$$df = (\#\ of\ rows - 1)(\#\ of\ columns - 1)$$
$$= (3\text{-}1)(2\text{-}1) = 2$$

```
CROSSTABS
  /TABLES=gender  BY party
  /STATISTIC=CHISQ.
```

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7.010[a]	2	.030
Likelihood Ratio	7.003	2	.030
Linear-by-Linear Association	6.758	1	.009
N of Valid Cases	980		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 49.35.

$$\chi^2(2) = 7.01, p = .03$$

- For any table larger than 2*2, use a modified ϕ, called Cramer's phi, ϕ_C, to measure the effect size of X^2

$$\phi_C = \sqrt{\frac{\chi^2}{N(L-1)}} = \sqrt{\frac{7.01}{980(2-1)}} = .08$$

Where L = min(# rows, # of columns)

$$\chi^2(2) = 7.01, p = .03, \phi_C = .08$$

- Reject null hypothesis and conclude that party affiliation varies by gender. But we cannot say how it varies – only that it varies. To make more specific claims, we have to conduct follow-up tests.

- Follow-up #1: Do males and females differ in their propensity to be either a Democrat or Republican?

	Party Affiliation	
Gender	Democrat	Republican
Females	279	225
Males	165	191

  - For this table, we omit the people who indicated that they were Independent

- Follow-up #2: Does the number of independents and non-independents differ for Males and Females?

	Party Affiliation	
		Non-Independent
Gender	Independent	(Democrat + Republican)
Females	73	504
Males	47	356

  - For this table, we combine Democrats and Republicans into a single column

- When we conduct these follow-up analyses, we will be able to make more focused conclusions.

Chapter 2: Appendix

A. Interesting and useful facts about the Chi-square distribution

- A squared standardized z-score is distributed $\chi^2(1)$

  If $Y \sim N(\mu,\sigma)$ then over repeated sampling $\frac{(y-\mu)^2}{\sigma^2} = \chi^2(1)$

- Suppose y_1 and y_2 are drawn independently from Y.

  Then $z_1^2 = \frac{(y_1-\mu)^2}{\sigma^2}$ and $z_2^2 = \frac{(y_2-\mu)^2}{\sigma^2}$

  and over repeated sampling, $z_1^2 + z_2^2 = \chi^2(2)$

- In general, for n independent observations from a normal population, the sum of the squared standardized values for the observations has a chi-square distribution with n degrees of freedom

  If $z_i^2 = \frac{(y_i-\mu)^2}{\sigma^2}$

  Then $\sum_{i=1}^{n} z_i^2 = \chi^2(n)$

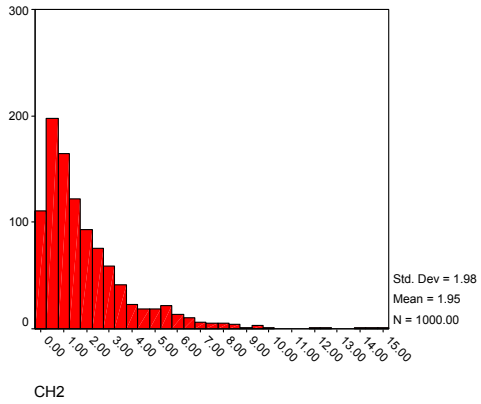
- If a random variable, Y_1, has a chi-square distribution with ν_1 degrees of freedom, and an independent random variable Y_2, has a chi-squared distribution with ν_2 degrees of freedom, then the new random variable formed from the sum of Y_1 and Y_2 has a chi-square distribution with $\nu_1 + \nu_2$ degrees of freedom.

  If $Y_1 \sim \chi^2(\nu_1)$ and $Y_2 \sim \chi^2(\nu_2)$

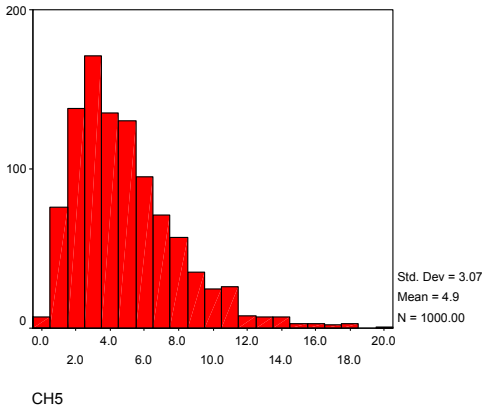
  Then $Y_1 + Y_2 \sim \chi^2(\nu_1 + \nu_2)$

  In other words, $\chi^2(\nu_1) + \chi^2(\nu_2) = \chi^2(\nu_1 + \nu_2)$

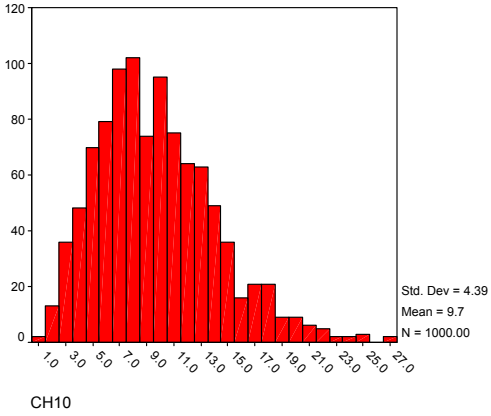
- Chi-Square *df*=2



Chi-Square *df*=5



- Chi-Square *df*=10



Chi-Square *df*=20