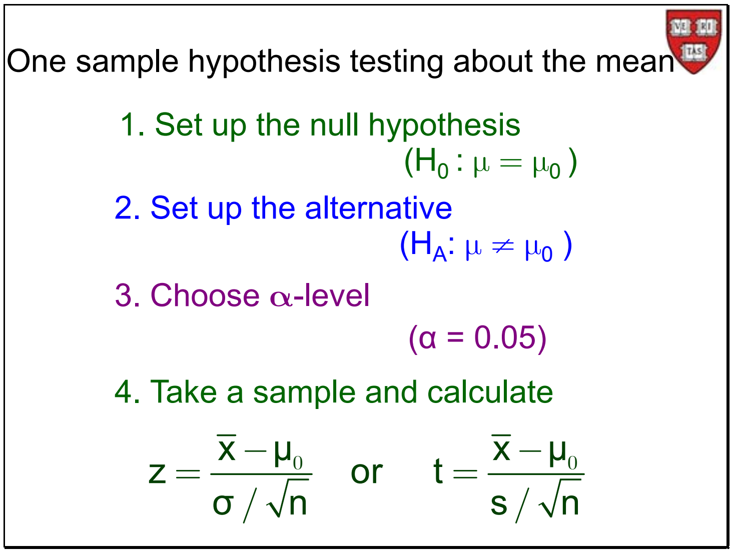
Marcello Pagano

[JOTTER 6 TWO-MEANS]

The testing of one mean is extended to two or more samples (ANOVA) with a fuller description of power.



One sample hypothesis testing about the mean

1. Set up the null hypothesis

   $(H_0 : \mu = \mu_0)$

2. Set up the alternative

   $(H_A: \mu \neq \mu_0)$

3. Choose α-level

   $(\alpha = 0.05)$

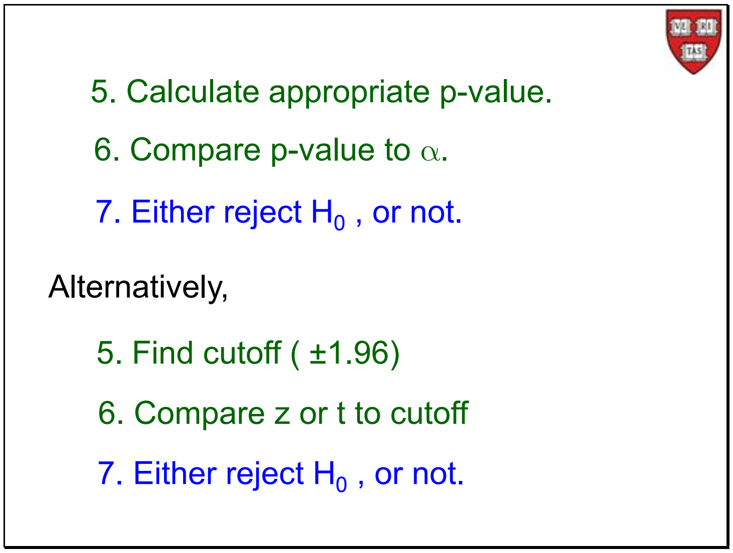
4. Take a sample and calculate

$$z = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}} \quad \text{or} \quad t = \frac{\overline{x} - \mu_0}{s / \sqrt{n}}$$

Let us quickly review the steps we take to test an hypothesis with a single sample: Step one, we set up the null hypothesis. For example, we can make a statement about the mean of the population; that it is equal to some μ_0. Step two, set up the complement of the null hypothesis. That is the alternative hypothesis; here we have the two-sided case. Step three, choose an alpha level; choose the ubiquitous 0.05.

Step four is to take a sample of size n and then calculate either Z or t, depending on whether we know σ, or not. In either case, what we are comparing the sample mean to the hypothesized population mean to see how far apart they are. If they are very close to each other, then we shall say that the data is consonant with the hypothesis. If they are far apart, then we argue that the data do not seem to be supporting the null hypothesis and so we reject the null hypothesis. The remaining point is how to judge what is small and what is big?

To this end, we divide by the appropriate standard deviation—we standardize; Z if we know σ, t otherwise. The Z, or the t, can now be compared to what we expect to see if the null hypothesis is true.



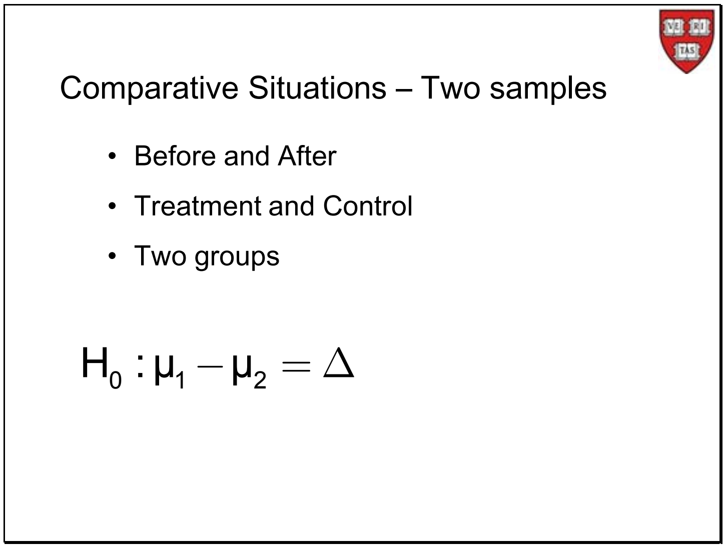
5. Calculate appropriate p-value.
6. Compare p-value to α.
7. Either reject H_0 , or not.

Alternatively,

5. Find cutoff (±1.96)
6. Compare z or t to cutoff
7. Either reject H_0 , or not.

We then compare these values to their appropriate cut-offs. So, for example, we can calculate the appropriate p-value for our statistic, compare that p-value to α, and either reject H_0 or not. If the p-value is less than α, we shall reject; if not, we shall not reject H_0.

Alternatively, we can look directly for the cut-off of the statistic. For example, if we were looking at Z our cut-off would be plus or minus 1.96. So reject H_0 if the magnitude of Z is bigger than 1.96; do not reject if it is less than 1.96 in magnitude. In the case of t, find the appropriate cut-off for the given degrees of freedom, and then proceed as with Z. In either case, we are going to come up with the decision to either reject the null hypothesis, or not.



Comparative Situations – Two samples

- Before and After
- Treatment and Control
- Two groups

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

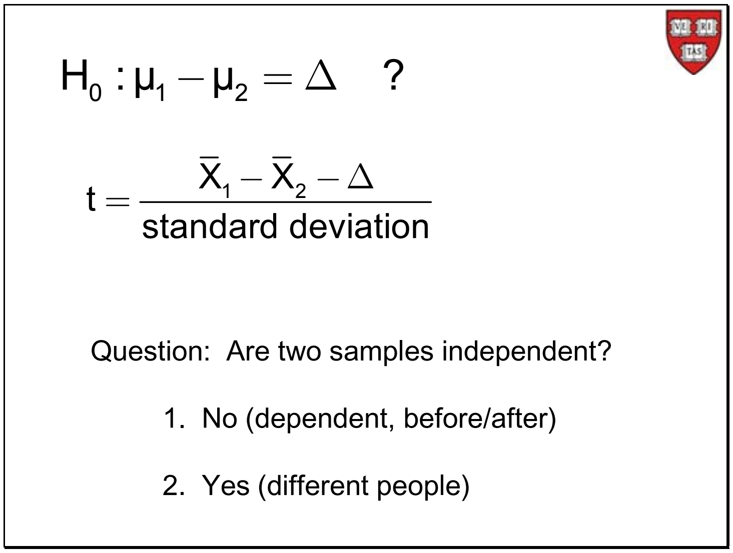
Now let us move on and generalize the situation to the case when we have two samples, and not just one. For example, if we have a before and after situation—weighing in before going on a diet and after going on a diet to judge whether the diet is any good.

Alternatively, we might have two groups of individuals: one group gets an experimental treatment and the other group, the control group, gets the standard treatment. Once again, it could be the same persons in the groups—for one week you give the experimental treatment and then you let that wear off, and then the next week you give that person a control treatment, for example.

Or you might have two groups of patients, 50 patients, get the treatment. Another, separate group of 50 patients, get the control. And that is a treatment and control situation.

Our null hypothesis here is going to deal with the difference in the two means, let us call it delta, and we are going to hypothesize something about delta. Now the most common hypothesis is that delta is equal to 0; that is that there is no difference, or no effect.

This is the generic way to set up the two sample problem.



$$H_0 : \mu_1 - \mu_2 = \Delta \quad ?$$

$$t = \frac{\bar{X}_1 - \bar{X}_2 - \Delta}{\text{standard deviation}}$$

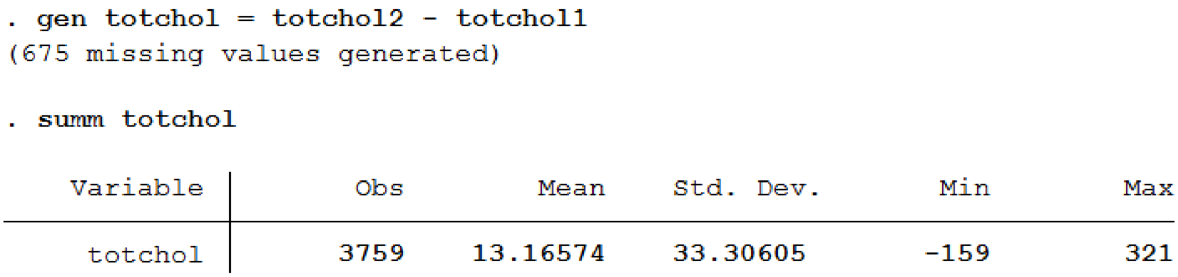
Question: Are two samples independent?

1. No (dependent, before/after)
2. Yes (different people)

As in the one-sample situation we need to determine our statistic. The thinking is very similar to the one-sample situation in that we look at the difference between the two sample means as our basic statistic. Also, as before, we compare this difference to the hypothesized difference, Δ, and then all that remains is to evaluate the size of this difference by dividing by an appropriate standard deviation.

The whole trick is going to be what we put in there for the standard deviation. As far as current theory has it, we need to classify our situation into one of two, depending on whether the two samples are independent, or not.

If they are not, for example, if we have a before and after situation on the same individual, or we are looking at the right eye of a rat and comparing it to the left eye of the same rat, then they are not independent. Then we perform one set of calculations. On the other hand, if the samples are independent— so we are talking about different people in the two groups—for example, we may be doing a male versus female, or possibly a young versus old comparison, or maybe we break up the population into two groups, one group gets one treatment and the other group gets the control treatment—then we perform a different set of calculations.



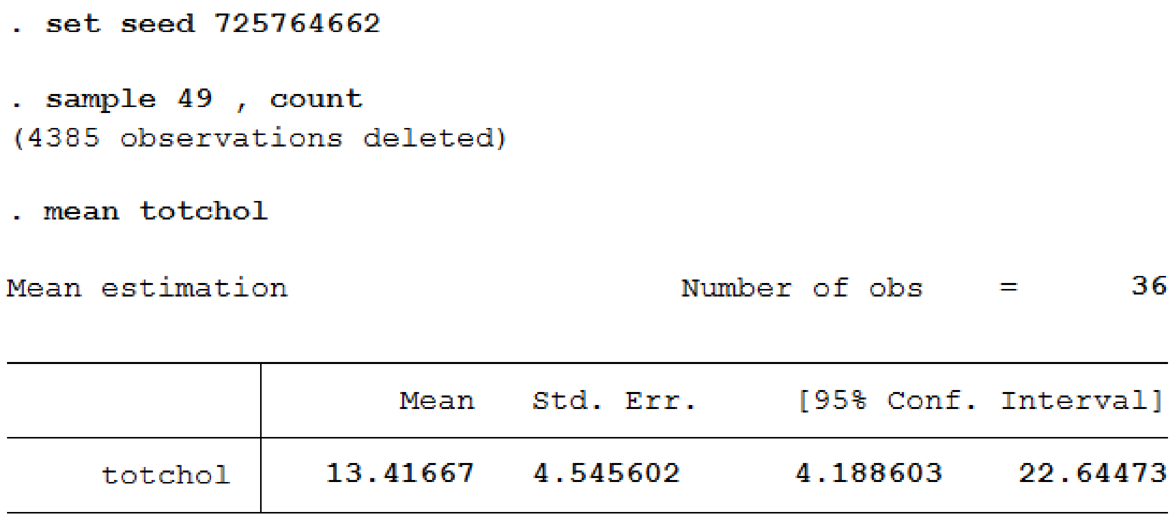
```
. gen totchol = totchol2 - totchol1
(675 missing values generated)

. summ totchol

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     totchol |      3759    13.16574    33.30605       -159        321
```

Let us first look at the dependent situation, and then we deal with the independent situation.

Suppose we are interested in what happens to the total cholesterol level between the first and second visit. We like the idea of each person serving as their own control, so we plan to measure each person twice, once at each visit.

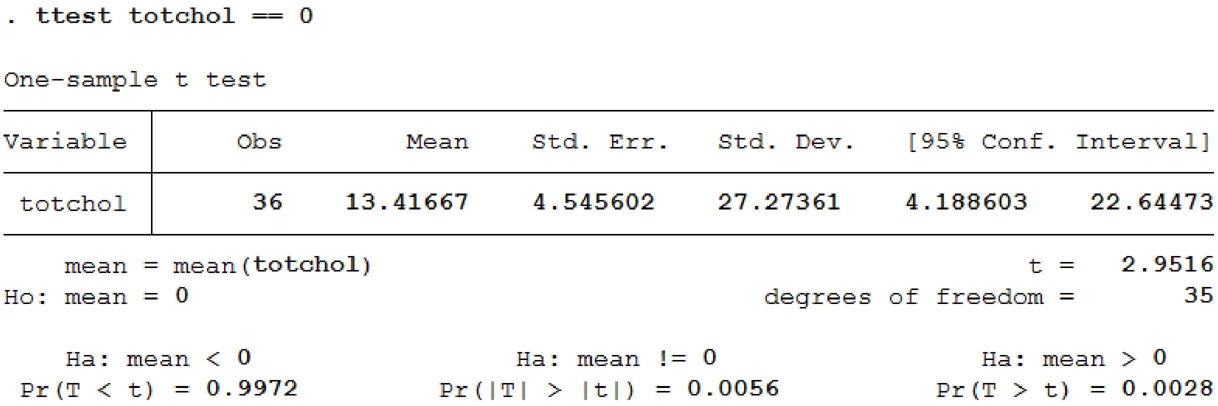
To analyze these data, generate a new variable, call it *totchol*, defined to be the difference between a person's cholesterol level at visit 2 and at visit 1. Once we do this, Stata returns the message that there are 675 missing values generated. This should be of concern to us, if this was a scientific enquiry, because what this is telling us is that there are 675 individuals who do not have both cholesterol readings at time 2 and at time 1. That might very well impact any sort of inference we want to make about the change in cholesterol level because these 675 who are missing might have a very different story to tell from the ones who are not missing.

Let us move on and not worry about this right now. Let us summarize *totchol* to find its mean is 13, there are 3,759 observations, and the standard deviation is 33, and not everybody shows an increase, so there are some negative numbers, but on average there is an increase in total cholesterol over the two visits.



```
. set seed 725764662

. sample 49 , count
(4385 observations deleted)

. mean totchol

Mean estimation                   Number of obs   =        36

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
     totchol |   13.41667   4.545602      4.188603    22.64473
--------------------------------------------------------------
```

Now let us take a sample of size 49 from our population. Now when we ask for the mean of this variable we find that the sample consists of only 36 of the 49 observations with *totchol* defined.[1]

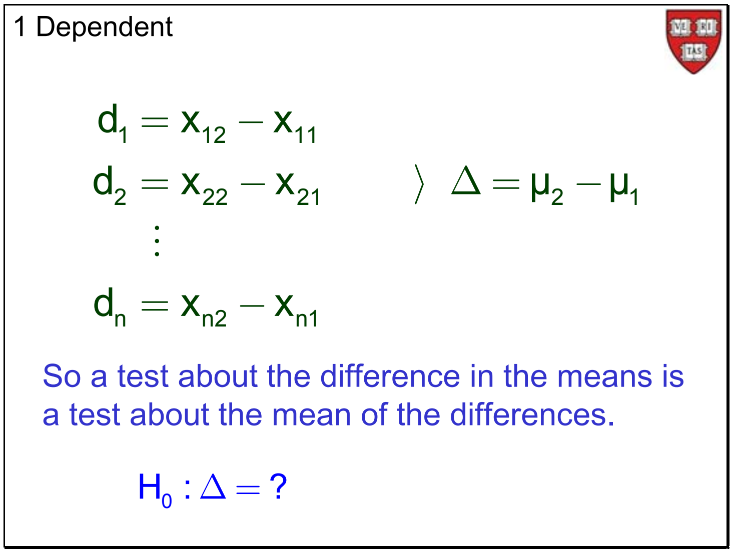


```
. ttest totchol == 0

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 totchol |      36    13.41667    4.545602    27.27361    4.188603    22.64473
------------------------------------------------------------------------------
    mean = mean(totchol)                                          t =   2.9516
Ho: mean = 0                                     degrees of freedom =       35

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.9972         Pr(|T| > |t|) = 0.0056          Pr(T > t) = 0.0028
```

Now we can test the null hypothesis that the difference in the means is zero. We see that the 95% confidence interval for the difference in cholesterol levels means between visit 2 and visit 1, is (4.2, 22.6). This does not include 0. So we know that the confidence interval approach to testing the null hypothesis, that the difference is 0, would reject that null hypothesis at the 5% level.

Now let's look at the hypothesis testing approach. The t statistic is 2.95 and there are 35 degrees of freedom, and if the alternative is that the mean is not 0, or the two-sided alternative, has a p-value attached to it of 0.0056. So we would reject the null hypothesis at the 5% level.

[1] This is different from the video because I did not set the seed in the video.



1 Dependent

$$d_1 = x_{12} - x_{11}$$
$$d_2 = x_{22} - x_{21} \qquad \rangle \quad \Delta = \mu_2 - \mu_1$$
$$\vdots$$
$$d_n = x_{n2} - x_{n1}$$

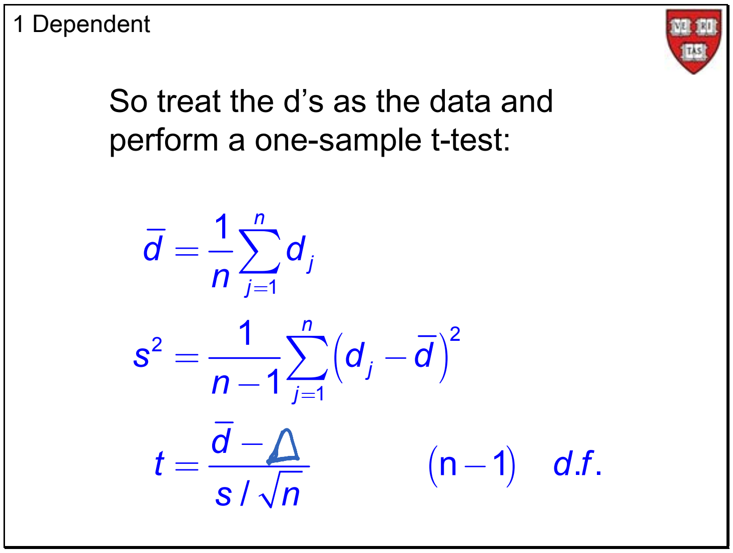
So a test about the difference in the means is a test about the mean of the differences.

$$H_0 : \Delta = ?$$

So in summary, in the dependent case look at each individual and take the difference between the value at time 2 and the value at time 1. Then ignore the individual x’s and concentrate on the individual differences.

We treat these differences just like we treated a single sample before, and call the mean of these differences, Δ; the difference between the two means.

So now we are back in the single sample situation with n observations, based on the n differences, and we can set up hypotheses about Δ—the most common hypothesis being H_0: Δ=0.



1 Dependent

So treat the d’s as the data and perform a one-sample t-test:

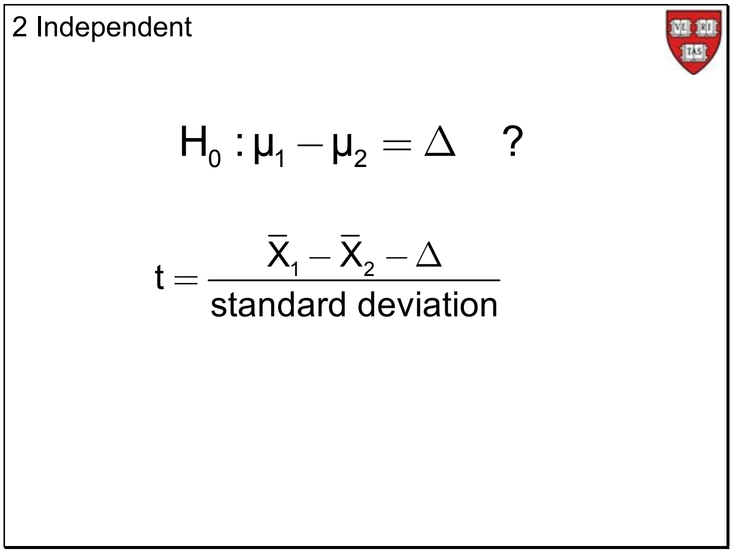
$$\bar{d} = \frac{1}{n}\sum_{j=1}^{n} d_j$$

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(d_j - \bar{d}\right)^2$$

$$t = \frac{\bar{d} - \Delta}{s / \sqrt{n}} \qquad (n-1) \quad d.f.$$

So we proceed exactly as before: Calculate the sample mean of the d’s; calculate the sample variance of the d's; and then take the mean of the d's, and divide by the standard error.

Insert the hypothesized value of Δ and under the null hypothesis, this statistic is distributed as a t with (n-1) degrees of freedom. And this is exactly what Stata reports.



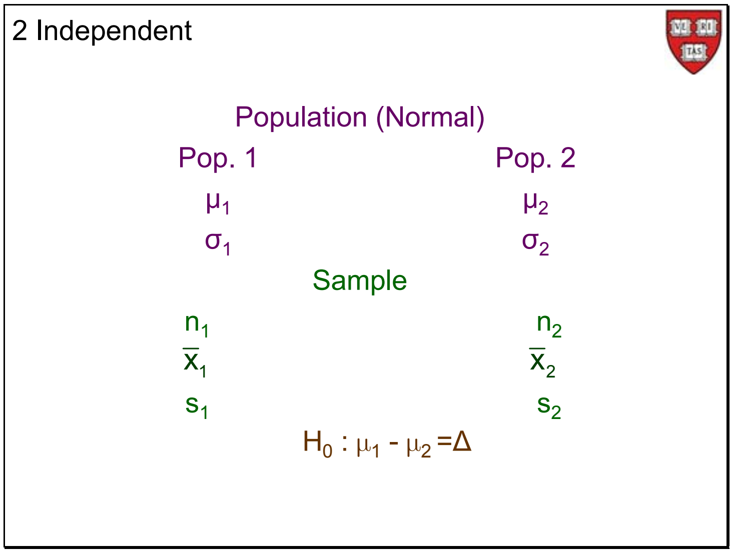
2 Independent

$$H_0 : \mu_1 - \mu_2 = \Delta \quad ?$$

$$t = \frac{\bar{X}_1 - \bar{X}_2 - \Delta}{\text{standard deviation}}$$

OK. So now we've got the dependent situation under our belts. Let's look at the independent situation. So we have two independent samples, and the hypothesis that we want to test is exactly the same as in the dependent case, namely a statement about the value of the difference between the two population-means; call it Δ.

And once again, the most common hypothesis is H_0: Δ=0. So there is no difference in these two groups. There is no difference between males and females. So this is our t just like before, the difference though is to be how we calculate the standard deviation.



2 Independent

Population (Normal)

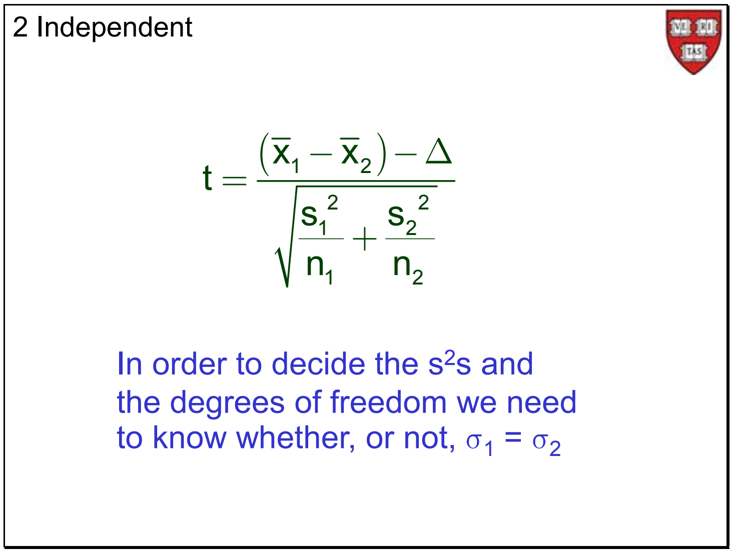
Pop. 1	Pop. 2
μ_1	μ_2
σ_1	σ_2

Sample

n_1	n_2
$\bar{x}_1$	$\bar{x}_2$
s_1	s_2

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

Let us establish some notation. We have two populations, and let us distinguish them by use of the subscripts 1 and 2. We take a sample of size n_1 from the first population and of size n_2 from the second population. Whereas, in the dependent case, these two sizes had to be the same, in this, independent case, they do not.



2 Independent

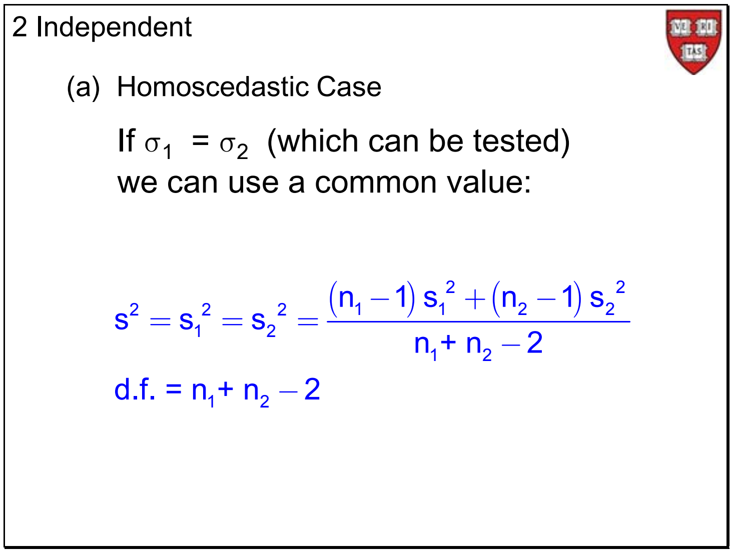
$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

In order to decide the s^2s and the degrees of freedom we need to know whether, or not, $\sigma_1 = \sigma_2$

One possible t that we could think of is to take the variance of the first sample divided by n_1 then the second variance and divide it by n_2, and take their sum and use that as our variance.

Unfortunately we are not done. Things get a little more complicated. We need to ask a second question. That second question is going to depend very much on what we know about the relative sizes of the two standard deviations—I am assuming we do not know the actual values of the two sigmas. What we need to know is if they are equal or not.

If the two population standard deviations are equal, then we call that the *homoscedastic* case; if not, we call that the *heteroscedastic* case. We can actually use our sample standard deviations to carry out a preliminary test to decide which of the two cases we are in, but that is not recommended.



2 Independent

(a) Homoscedastic Case

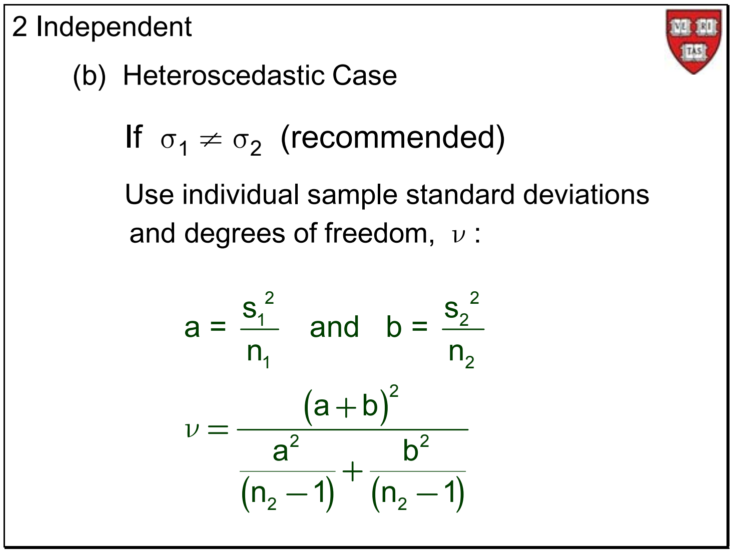
If $\sigma_1 = \sigma_2$ (which can be tested) we can use a common value:

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

d.f. = $n_1 + n_2 - 2$

In the homoscedastic case, we create a single estimator of the common variance, displayed above. Plug that into the t-statistic and it is distributed with $n_1 + n_2 - 2$ degrees of freedom.

Now here we see what price we paid when we had the dependent case. In the dependent case we had $n_1 - 1$ degrees of freedom, whereas here we have twice as many (assuming equal sample sizes). The degrees of freedom are related to the sample size(s), so the more degrees of freedom we have, the better it is.



2 Independent

(b) Heteroscedastic Case

If $\sigma_1 \neq \sigma_2$ (recommended)

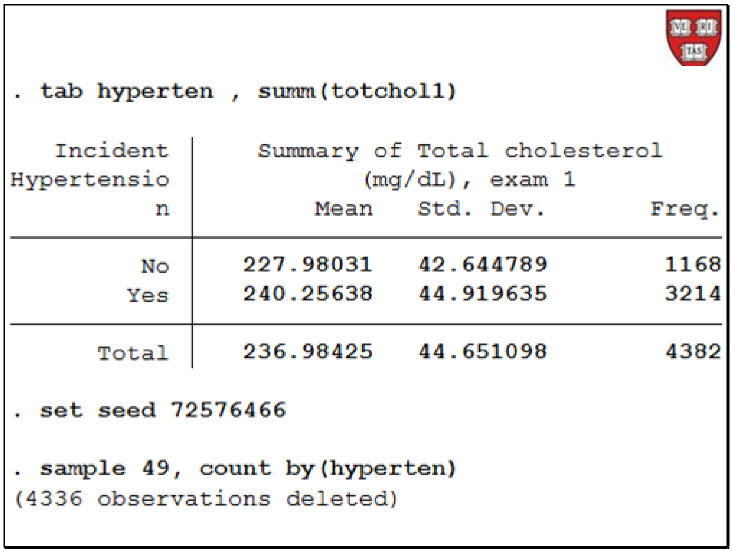
Use individual sample standard deviations and degrees of freedom, ν :

$$a = \frac{s_1^{\,2}}{n_1} \quad \text{and} \quad b = \frac{s_2^{\,2}}{n_2}$$

$$\nu = \frac{(a+b)^2}{\dfrac{a^2}{(n_2 - 1)} + \dfrac{b^2}{(n_2 - 1)}}$$

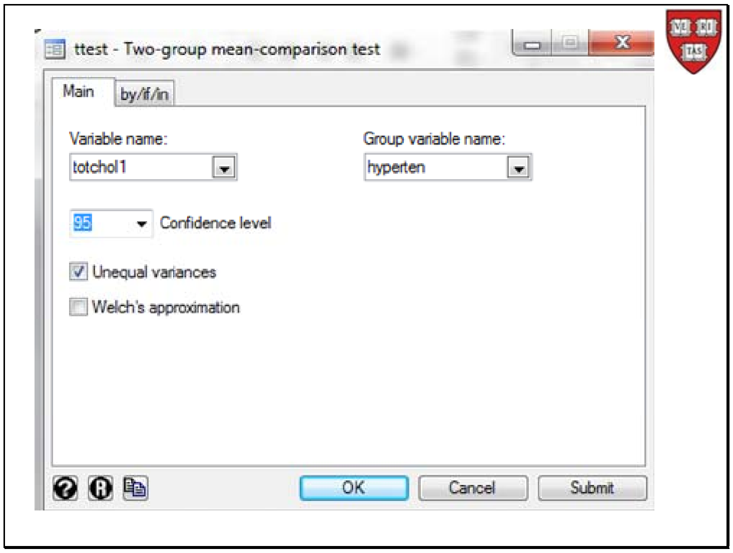
Now let us look at the heteroscedastic case. This is the way I would recommend you proceed in general; namely, act as if you are in the heteroscedastic case. You lose a few degrees of freedom, but for those you buy a little protection from an assumption (equality of the standard deviations) you do not have to make, and even if true, should not cause problems.

The complications with the heteroscedastic case are: (i) the degrees of freedom are a little bit more complex (for Stata) to calculate; and, (ii) they are not exact, but provide an approximation. The one you see above is one approximation; Stata provides us with a choice.



```
. tab hyperten , summ(totchol1)

  Incident  |   Summary of Total cholesterol
Hypertensio |         (mg/dL), exam 1
          n |        Mean   Std. Dev.       Freq.
------------+------------------------------------
         No |   227.98031   42.644789        1168
        Yes |   240.25638   44.919635        3214
------------+------------------------------------
      Total |   236.98425   44.651098        4382

. set seed 72576466

. sample 49, count by(hyperten)
(4336 observations deleted)
```

So let us go to Stata and see how easy it is to actually do these calculations. First, let us set the problem up: If we split the population into two groups, those who are not hypertensive at the beginning (*hyperten* is no, N_2 = 1168) and those who are (N_1 = 3214), and for whom we have total cholesterol level readings at visit 1. For those who are not hypertensive coming in, their mean total cholesterol is about 228, and those who did have hypertension coming in, their cholesterol level is about 240. So indeed there is a small difference in total cholesterol between these two groups. Let us proceed to take samples, of size 49, from each of these groups and see if we can detect this difference based on inference from our samples.



ttest - Two-group mean-comparison test

Variable name: totchol1

Group variable name: hyperten

Confidence level: 95

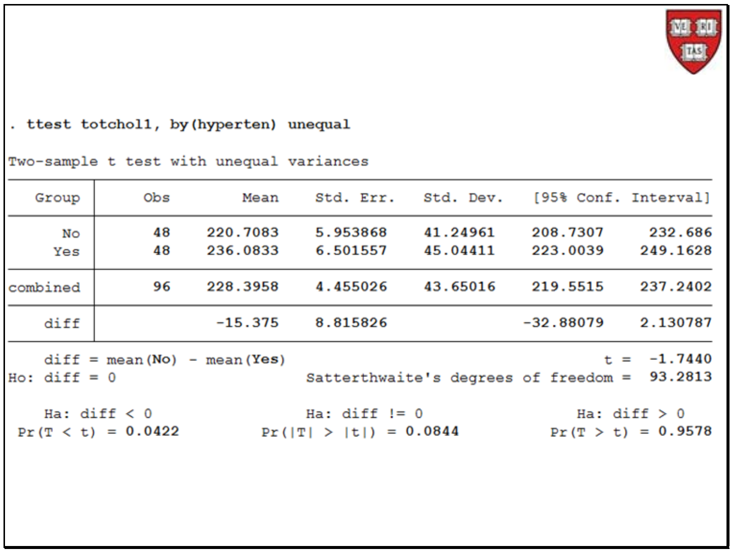
☑ Unequal variances

□ Welch's approximation

Now we go to Statistics > Summary tables > Classical tests of hypotheses, and what we want is the > Two-group mean comparison test. So there it is.

The variable that we want to test is going to be our total cholesterol level at time 1, so it's *totchol1*. And the Group variable name is *hyperten*. As mentioned above, I recommend you choose the "Unequal variances" option.

We could check the Welch approximation box and get a different approximation, but let us first leave it at the default and get what is called the Satterthwaite approximation. I leave it up to you to check the Manual to see investigate the details about the differences between these two, if you are interested. By all means rerun your analysis using the one you did not previously use and see if you get different results. They usually give very similar answers.



```
. ttest totchol1, by(hyperten) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      No |      48    220.7083    5.953868    41.24961    208.7307     232.686
     Yes |      48    236.0833    6.501557    45.04411    223.0039    249.1628
---------+--------------------------------------------------------------------
combined |      96    228.3958    4.455026    43.65016    219.5515    237.2402
---------+--------------------------------------------------------------------
    diff |             -15.375    8.815826               -32.88079    2.130787
------------------------------------------------------------------------------
    diff = mean(No) - mean(Yes)                                   t =  -1.7440
Ho: diff = 0                         Satterthwaite's degrees of freedom =  93.2813

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0422         Pr(|T| > |t|) = 0.0844          Pr(T > t) = 0.9578
```

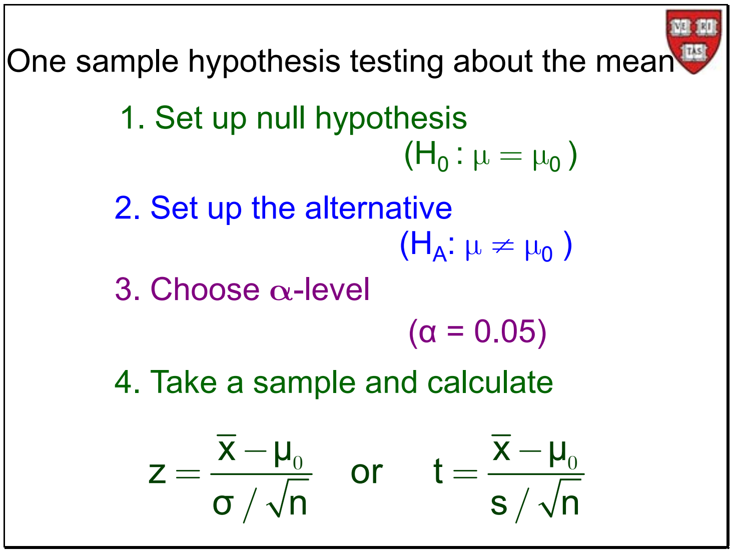
The first thing we notice on the output is that we lost two observations, one from each sample, because of missing data. So of the 48 who do not have hypertension, their mean is 220. And for the 48 with hypertension, their mean was 236. These means are lower than in the population, but the standard deviations are pretty close to what they are in the population.

The differences are reported in the "diff" row. The difference in the sample means is 220.7083-236.0833 = -15.375. Those differences have a reported standard error of 8.81. And if we want to use the confidence interval approach to test the hypothesis of any differences, we would find that the 95% confidence interval is (- 32.88, 2.13) which includes the value 0. So by using the confidence interval approach, we would not reject the null hypothesis that the two means are the same.

If we perform a hypothesis test we see at the bottom that the p-value associated with a two sided test is 0.0844, and so we would not reject the null hypothesis of equality of the two group means.

Explore what happens with the Welch approximation and also explore what happens had you made the homoscedastic assumption.

Type 2 Error – Power



One sample hypothesis testing about the mean

1. Set up null hypothesis

   $(H_0 : \mu = \mu_0)$

2. Set up the alternative

   $(H_A : \mu \neq \mu_0)$

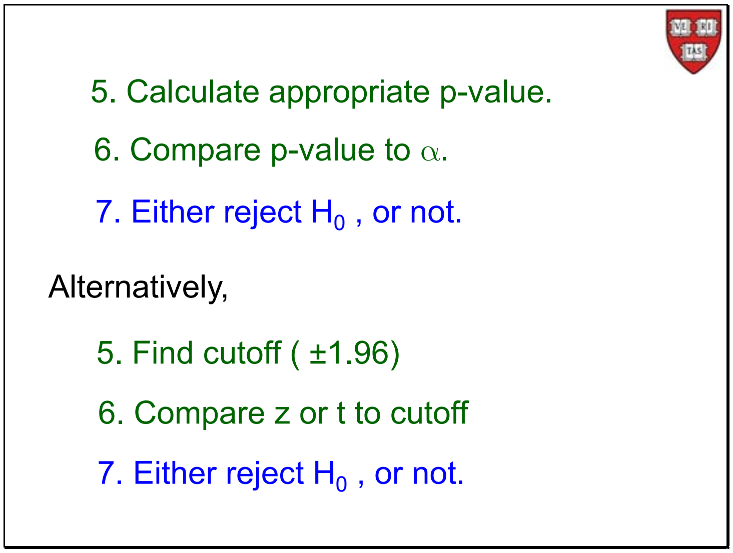
3. Choose α-level

   $(\alpha = 0.05)$

4. Take a sample and calculate

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \quad \text{or} \quad t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Here is an outline of what we have done up to now to test a hypothesis about the mean of a population. Here are the first four steps we take.

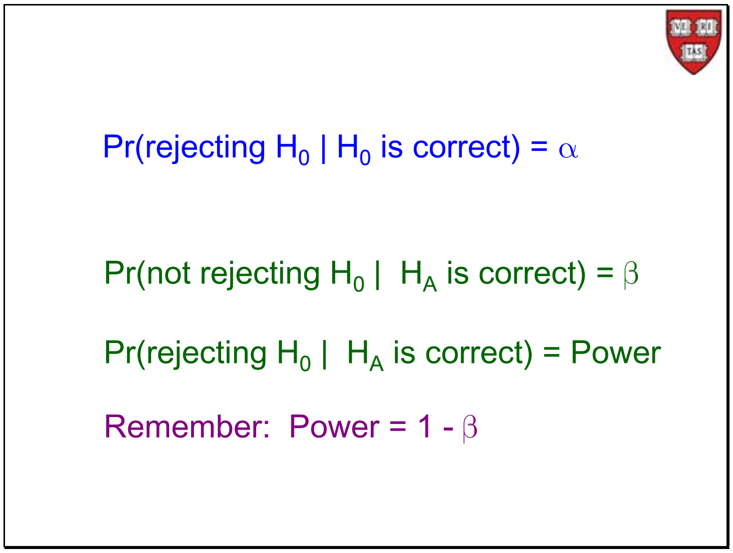


5. Calculate appropriate p-value.
6. Compare p-value to α.
7. Either reject H_0 , or not.

Alternatively,

5. Find cutoff (±1.96)
6. Compare z or t to cutoff
7. Either reject H_0 , or not.

Then we take the fifth step, sixth, and seventh steps whichever way we go, whether we calculate the appropriate p-value or whether we do it by using our cutoffs for the appropriate alpha.



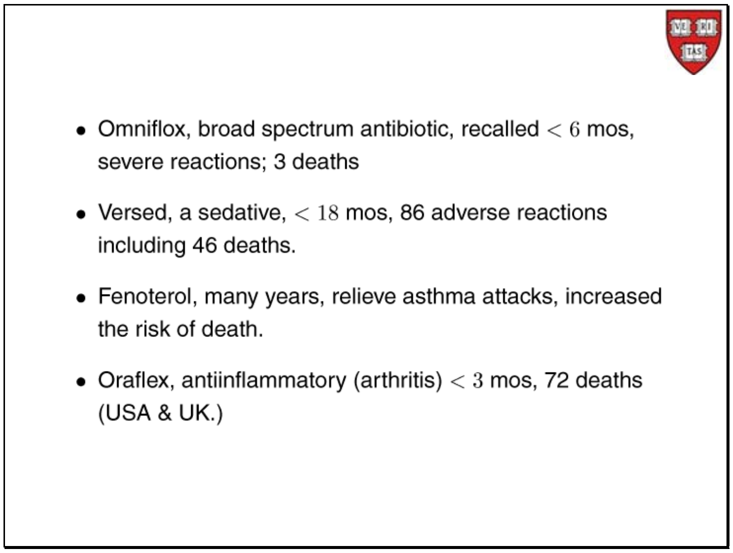
Pr(rejecting H_0 | H_0 is correct) = α

Pr(not rejecting H_0 | H_A is correct) = β

Pr(rejecting H_0 | H_A is correct) = Power

Remember: Power = 1 - β

What we have not paid attention to up to now is what happens to our power, or the type II error. We have set the alpha at 0.05, and not mentioned beta. So what can be said about the probability of not rejecting H0 when we should be rejecting it? Or, if we like to be positive in life, we can talk about 1 minus beta, namely the power.



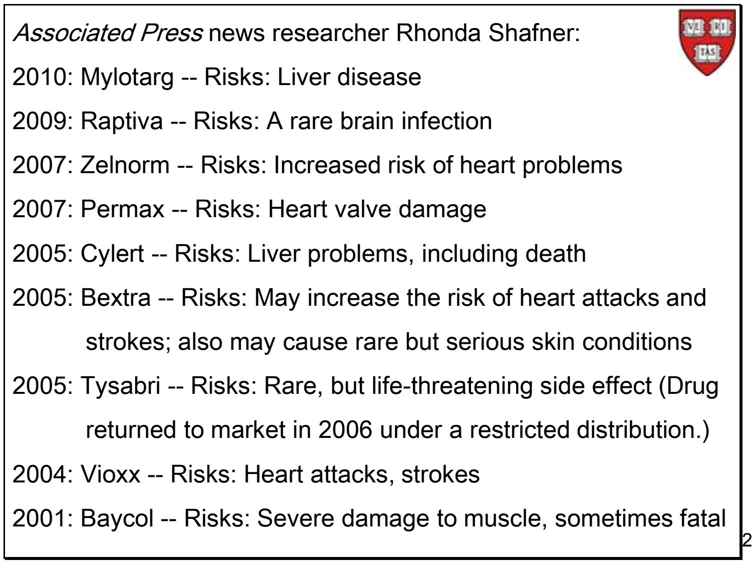
- Omniflox, broad spectrum antibiotic, recalled < 6 mos, severe reactions; 3 deaths
- Versed, a sedative, < 18 mos, 86 adverse reactions including 46 deaths.
- Fenoterol, many years, relieve asthma attacks, increased the risk of death.
- Oraflex, antiinflammatory (arthritis) < 3 mos, 72 deaths (USA & UK.)

Is this something we should be concerned about in real life? The very dramatic example of when this happens is in clinical trials, which usually form the knowledge base for the FDA to decide whether certain drugs should be allowed out onto the market.

An important aspect of these clinical trials is to study the drugs' side effects. Here are some drugs that were deemed safe enough to be allowed out onto the market, but were subsequently recalled because of their safety profiles. Possibly the null hypothesis of not having serious drug effects was not rejected.

The first one, Omniflox, was recalled less than six months after it had been released. There were all sorts of severe reactions associated with it—it was a broad spectrum antibiotic—including three deaths. Versed, was out for about 18 months, and had 46 deaths associated with it. And so on. So the question is, can we avoid this? It seems like we cannot, even though the FDA has a terrific track record.

It seems like risks are inevitable, as we cannot cover all possibilities. For example, suppose a negative effect takes six years to manifest itself. If the clinical trial lasts five years, you are not going to see it. This is an argument for for post-marketing monitoring to keep an eye on what is going on.



Associated Press news researcher Rhonda Shafner:

2010: Mylotarg -- Risks: Liver disease

2009: Raptiva -- Risks: A rare brain infection

2007: Zelnorm -- Risks: Increased risk of heart problems

2007: Permax -- Risks: Heart valve damage

2005: Cylert -- Risks: Liver problems, including death

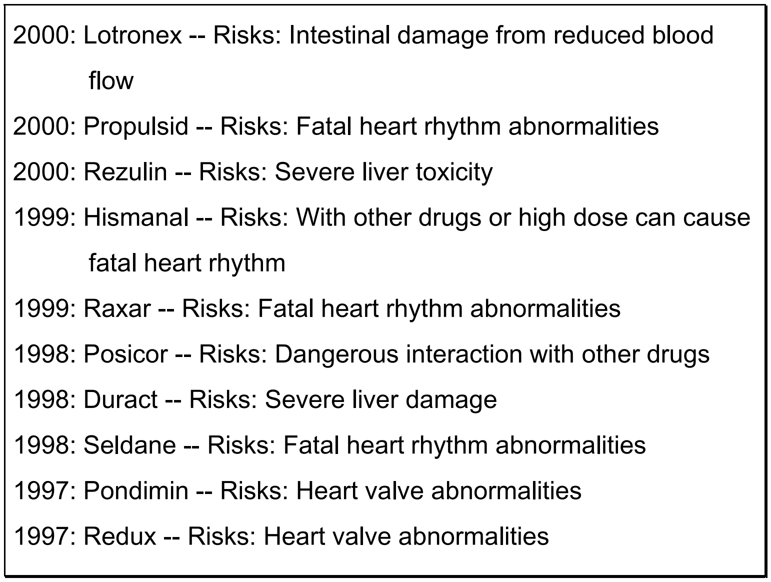
2005: Bextra -- Risks: May increase the risk of heart attacks and strokes; also may cause rare but serious skin conditions

2005: Tysabri -- Risks: Rare, but life-threatening side effect (Drug returned to market in 2006 under a restricted distribution.)

2004: Vioxx -- Risks: Heart attacks, strokes

2001: Baycol -- Risks: Severe damage to muscle, sometimes fatal[2]

[2] http://www.drug-injury.com/druginjurycom/2010/07/unsafe-drug-recall-decision-determination-factors-medicines-withdrawn-us-fda-history.html



2000: Lotronex -- Risks: Intestinal damage from reduced blood flow

2000: Propulsid -- Risks: Fatal heart rhythm abnormalities

2000: Rezulin -- Risks: Severe liver toxicity

1999: Hismanal -- Risks: With other drugs or high dose can cause fatal heart rhythm

1999: Raxar -- Risks: Fatal heart rhythm abnormalities

1998: Posicor -- Risks: Dangerous interaction with other drugs

1998: Duract -- Risks: Severe liver damage

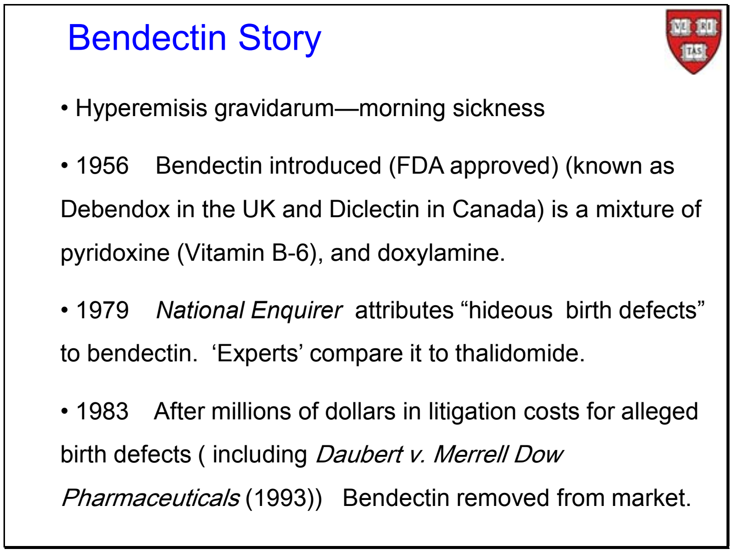
1998: Seldane -- Risks: Fatal heart rhythm abnormalities

1997: Pondimin -- Risks: Heart valve abnormalities

1997: Redux -- Risks: Heart valve abnormalities

These recalled drugs are historical recalls, but the problem persists. Here is a list of more recent recalls compiled by Rhonda Shafner of the Associated Press, that covers the last fourteen years or so.

Lest we go away with the wrong impression, the opposite can happen too, namely a good drug taken off the market for the wrong reasons.

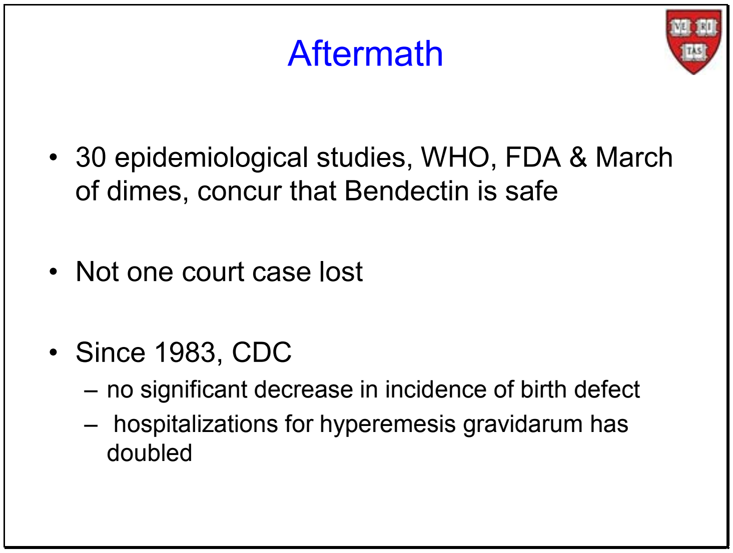


Bendectin Story

- Hyperemisis gravidarum—morning sickness
- 1956 Bendectin introduced (FDA approved) (known as Debendox in the UK and Diclectin in Canada) is a mixture of pyridoxine (Vitamin B-6), and doxylamine.
- 1979 *National Enquirer* attributes "hideous birth defects" to bendectin. 'Experts' compare it to thalidomide.
- 1983 After millions of dollars in litigation costs for alleged birth defects (including *Daubert v. Merrell Dow Pharmaceuticals* (1993)) Bendectin removed from market.

Let me tell you the Bendectin story. Bendectin was a drug that was introduced in 1956 to combat morning sickness (NVP—nausea and vomiting of pregnancy). So for pregnant women suffering from morning sickness, they had this drug, Bendectin. In the UK and Canada it has a different name. It is available all across the world except in the US. And when we ask, why, one discovers that in 1979 the National Enquirer published an article about Bendectin. In the article they had phrases like "hideous birth defects" are associated with Bendectin. They had statements by “experts” that compared Bendectin to thalidomide. And back in those days, the word thalidomide was—it still is—a very scary thought for a pregnant woman. But back in those days, it raised all sorts of horrible mental pictures.

For those of you who do not know what the National Enquirer is, I do not know how to explain it. It is the sort of publication you see when you are waiting in line in the supermarket to check out your purchases and you see those—I do not want to use the word rag but—kinds of newspapers that you really should not be spending any money buying but you read while you are waiting in line for a good laugh.



Aftermath

- 30 epidemiological studies, WHO, FDA & March of dimes, concur that Bendectin is safe
- Not one court case lost
- Since 1983, CDC
  - no significant decrease in incidence of birth defect
  - hospitalizations for hyperemesis gravidarum has doubled

But what happened subsequent to that story, by 1983 after millions of dollars in litigation costs for alleged birth defects, Bendectin was removed from the market by the company that manufactured it. Now it is sad because at that time that was the only drug that had been approved by the FDA to handle morning sickness. In its defense, there were 30 large epidemiological studies done all over the world by very reputable organizations such as the WHO, the World Health Organization, the FDA, the March of Dimes.

They all came to the same conclusion, namely that Bendectin is safe.

So you had on the one side the National Enquirer claiming something, and then you had the preponderance of scientific evidence and the scientific community saying that what was published was nonsense. Indeed, not one court case was lost, but the manufacturer did not wish to take any further risks. So sometimes it is not science that dictates on scientific issues.

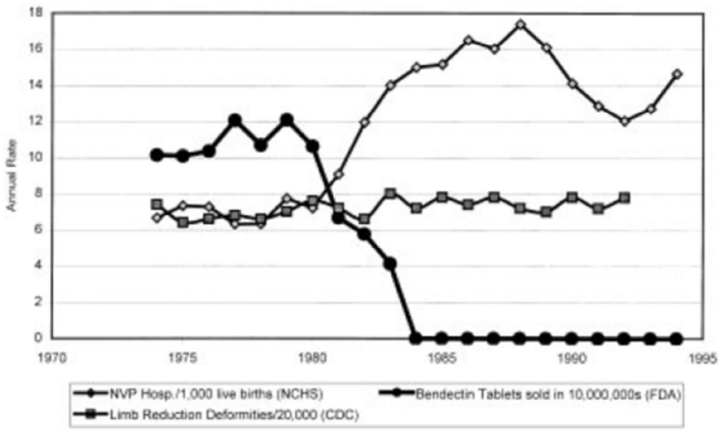


Figure 12. Public health data related to Bendectin therapy.

JS Kutcher, A Engle, J Firth, SH. Lamm ,
Birth Defects Research (Part A) 67:88–97 (2003)

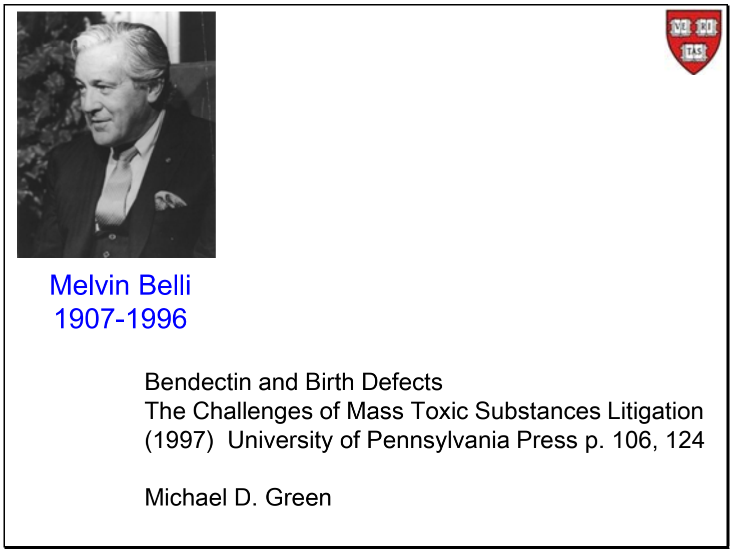
Some twenty years after the withdrawal of Bendectin from the market, still with no replacement drug, a study was done to shed some light on the issue. It is an ecological study, namely done at the country level and not at the individual level, but it is interesting nonetheless. The argument follows along the line: if you have a period when Bendectin is on the market and it's supposed to be doing some harm, then when you take it off the market the frequency of that harm should decrease. Look at the above graph of three lines.

The solid black line that starts at the top on the left and ends at the bottom on the right represents Bendectin sales in the US. It drops precipitously in the early 80s to reflect the fact that it was taken off the market. Prior to being taken off the market, the sales had been substantial.

Now remember that the drug was being blamed for causing birth defects. The middle line is the number of birth defects in the country—about as level a line as you can imagine. Taking the drug off the market does not seem to have had any impact on this line.

What about the poor pregnant women suffering from morning sickness? Focus on the third line (NVP) in the graph, the one that almost perfectly mirrors the Bendectin sales line. The line represents hospitalizations for the side effects of morning sickness, standardized by the number of births.

Clearly, this is an ecological study so all sorts of other factors may be influencing what is going on in this graph, but we are not just observing it is not a single relationship, that could be easily explained away, we are seeing two patterns: (i) it seems that taking Bendectin off the market was not followed by any decrease in birth defects, and (ii) the number of hospitalizations, for precisely the effects Bendectin was supposed to alleviate, increased after Bendectin was taken off the market.

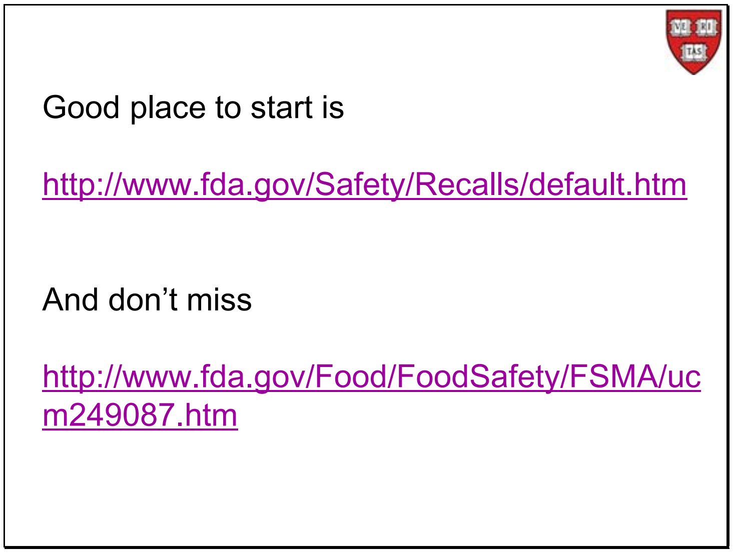


Melvin Belli
1907-1996

Bendectin and Birth Defects
The Challenges of Mass Toxic Substances Litigation
(1997) University of Pennsylvania Press p. 106, 124

Michael D. Green

The other question is why did the National Enquirer do this in the first place? Why did they publish the article? Well it turns out that they were fed the story by Melvin Belli, a lawyer, who had an interest in these mass tort cases and was suing the manufacturers of Bendectin and stood to make a lot of money had they been found culpable.



Good place to start is

http://www.fda.gov/Safety/Recalls/default.htm

And don't miss

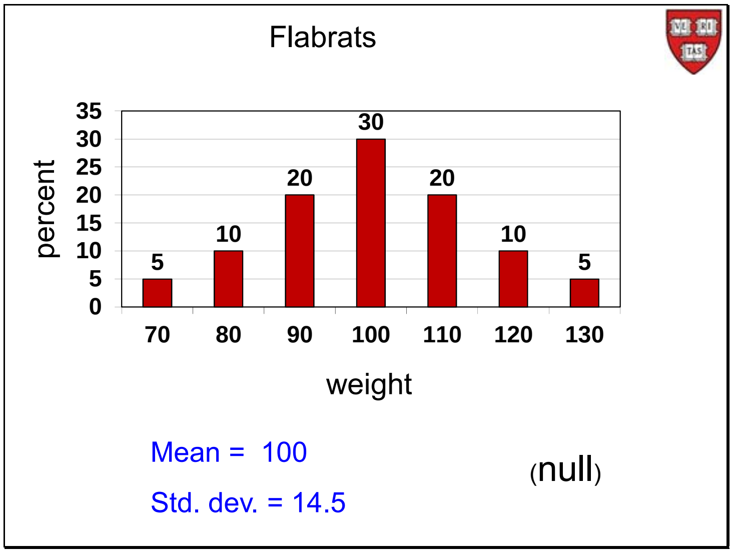
http://www.fda.gov/Food/FoodSafety/FSMA/ucm249087.htm

If you are interested in this topic, a good place to start is the FDA (Food and Drug Administration). They track recalls, and they do it for medications, for medical devices, et cetera. A definite do-not-miss is the bottom URL. It is a little difficult sometimes to find what you are really looking for in these large government agencies, but it is all there if you look carefully.

Power

Flabrat

The next topic we want to investigate is power. And to help us explain power, we turn to our friend the flabrat. Some years ago, when I first talked about flabrats, a student gave me this picture and said, here is a flabrat. This flabrat is eating a little bit of leaf here. And it's not really a flabrat, but anyway returning to statistics, flabrats are fabulous because they are sensitive to their diets and their weight. So we use them in the lab and just measure their weights.



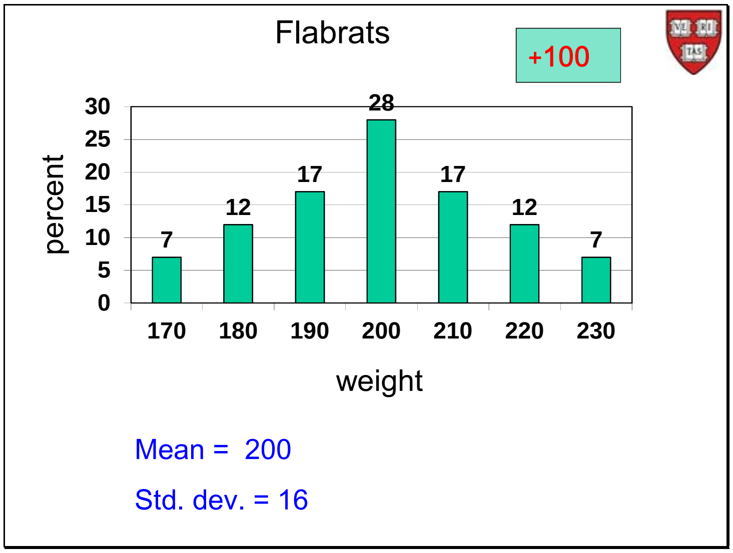
Flabrats

Mean = 100

Std. dev. = 14.5

(null)

Here is a bar chart describing the weights of flabrats out in the wild. Not that they are very wild, but we see that 5% weigh 70 units, 10% weigh 80 units, 20% weigh 90 units, and so on. If you need a unit of weight, you can think of your favorite unit, such as a gram if you want to, but it does not have to be, and if we put this distribution into our statistical calculator, we find out that its mean is 100 and its standard deviation is 14.5. Let us call this our null distribution.



Mean = 200

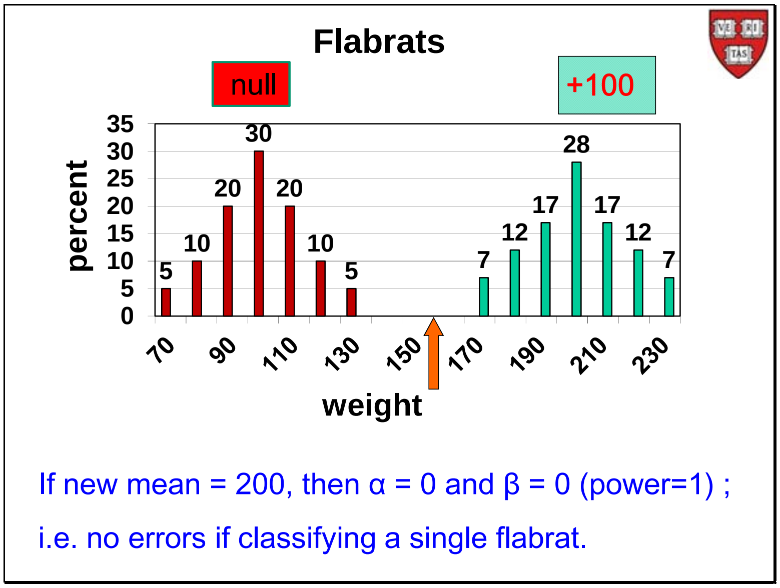
Std. dev. = 16

If we feed our flabrats a diet called the “Plus 100” diet, we shift the whole distribution 100 units to the right and the mean becomes 200, but the standard deviation also changed a bit. Instead of having 5 on the extremities, we have 7% on the extremities, amongst other changes. Now the standard deviation is 16. (This was done so as not to confuse you later.)

Here is the problem: we have two labs set up on either side of a corridor, and the lab on the left houses the flabrats with their natural diet (null), and on the lab on the right houses the flabrats who are on the “Plus 100” diet.

One morning you come to work and there is a flabrat in the corridor, and you say to yourself, oh my gosh, where did this flabrat come from? Does this flabrat belong in the left lab or the right lab?

The challenge is to properly classify the wandering flabrat.



If new mean = 200, then α = 0 and β = 0 (power=1) ;

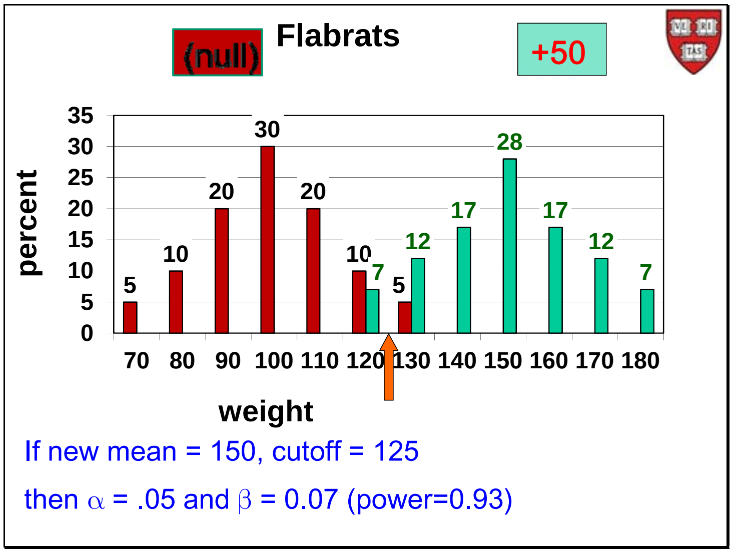
i.e. no errors if classifying a single flabrat.

If we put the two distributions side by side, how would you decide to classify the wandering flabrat? Surely, the decision should be made on how much the wandering flabrat weighs. So you can pick it up, weigh it, and then decide.

Then a sensible criterion might be, choose a cutoff and a flabrat weighing less than the cutoff will be sent to the left lab, the one housing the null distribution, and a flabrat weighing more than the cutoff will be sent to the right lab, the one housing the “Plus 100”.

Placing the cutoff to the right of 130 and to the left of 170 will give us perfect discrimination because the two curves do not overlap.

If we place this in a hypothesis testing framework, we see that both our alpha and beta are going to be zero, and thus our power is one.



If new mean = 150, cutoff = 125

then α = .05 and β = 0.07 (power=0.93)

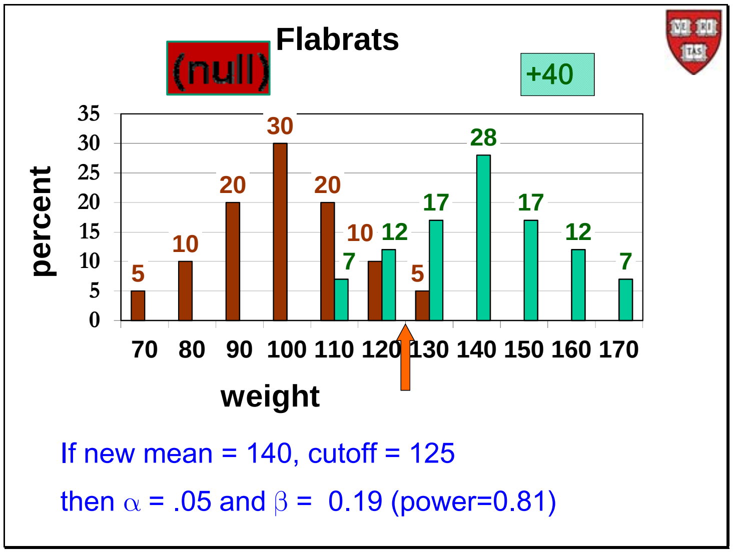
Your friend, one floor up, is also studying flabrats, except she is using a “Plus 50” diet as her alternative diet.

If we place the “Plus 50” distribution next to the null, we see overlap: there are flabrats in both groups who weigh 120 units (10% of the null and 7% of the “Plus 50”s) and 130 units (5% of the null and 12% of the “Plus 50”s). So now our discrimination is not as clean, and we may have some confusion—for example, if our wandering flabrat weighs 120 or 130 units.

But it still makes sense, since the “Plus 50”s tend to weigh more than the nulls, to use a weight cutoff as our discriminator, even though some of the nulls weigh more than some of the “Plus 50”s. And that is the source of our potential errors.

Suppose we want to keep our alpha error at most 0.05, then that means the cutoff must be to the right of 120. If we make 125 our cutoff, then alpha is 0.05 and beta is automatically 0.07 (and so the power is 93%). Anything much bigger, let us say, more than 130, will make our alpha smaller but it will increase our beta, and thus decrease our power.

So, in general, shifting the cutoff to the left increases alpha and decreases beta (increases power), with the opposite effect if we shift the cutoff to the right (decrease alpha, increase beta, decrease power). This is exactly what happens with hypothesis testing—we can associate the null with the flabrat coming from the null distribution, and the alternative with the flabrat coming from the “Plus 50” distribution. In this case we say we have 93% power to distinguish between the null and the “Plus 50” distribution on the basis of a single observation.

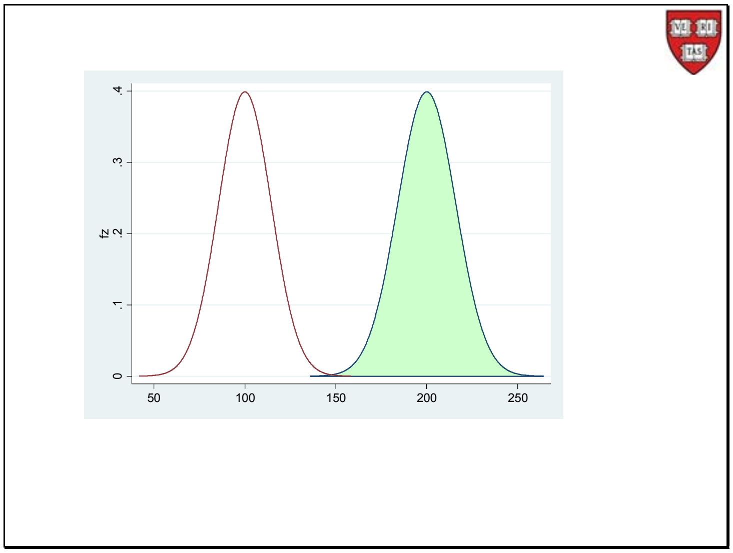


If new mean = 140, cutoff = 125

then α = .05 and β = 0.19 (power=0.81)

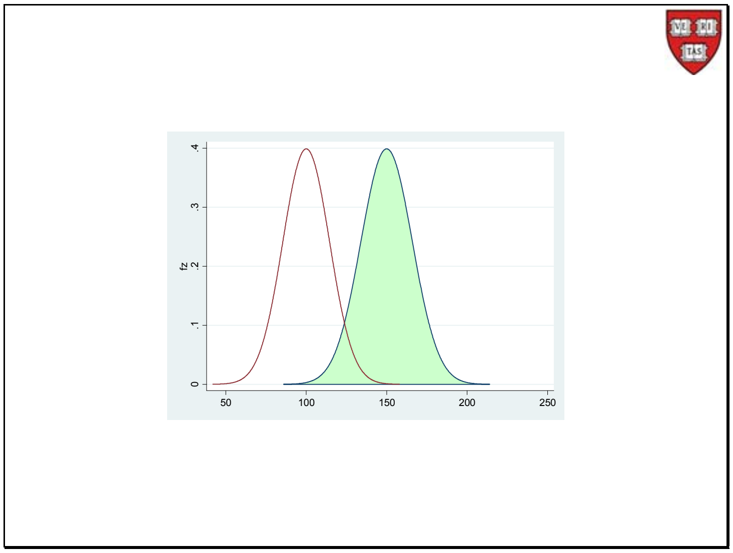
Your friend, two floors up, is also studying flabrats but with the subtler “Plus 40” diet. Now we see a bigger overlap.

Suppose that to retain the alpha at 5% you keep your cut-off at 125, then what does this do to the power (and beta)? Because of the bigger overlap we see that not only do we have the 7% falling at 110, but we also have 12% falling at 120, and thus both groups weigh less than the cutoff of 125, and thus a wandering flabrat from either of these groups would be classified as being in the null. So our beta now, because of the larger overlap, has increased to 0.19, dragging the power down to 81%.

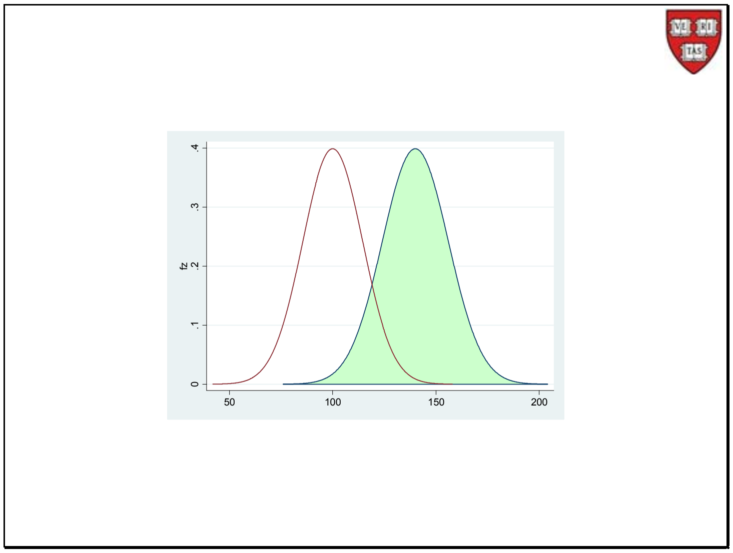


To smooth things out, let us bid farewell to the flabrats and look at smooth densities instead of our barcharts. It is much easier to now slide the alternative up and down the horizontal scale.

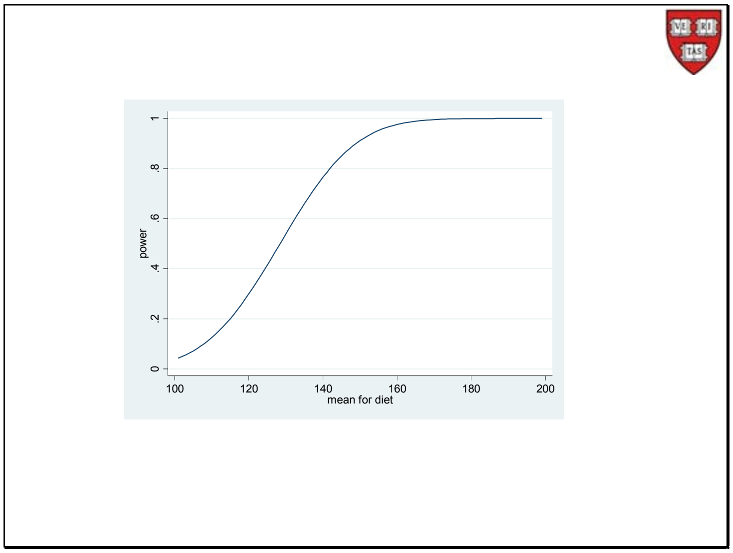
Above you see the analog of the “Plus 100” diet. Since these are two normal curves they both theoretically go off to infinity at both ends, so even at “Plus 100” there will be a tiny bit of an overlap of the two curves, and neither alpha nor beta ever is perfectly zero.



For the “Plus 50”, above, the overlap between the two distributions is much more noticeable than it was for the “Plus 100” situation, above. Now the mean is at 150, closer ot the null mean of 100, than when the mean was at 200 for the “plus 100”.



This is the chart for the “Plus 40” diet; much greater overlap because now the mean for the alternative is at 140.

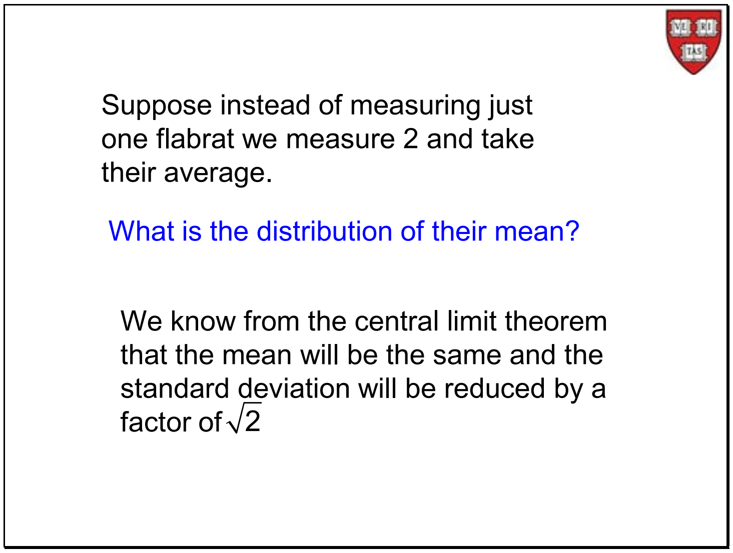


We can encapsulate this changing alternative with this curve, called the *power curve,* which displays the probability of rejecting the null hypothesis as a function of how far away the mean of the alternative is from the null mean. So you can see if there is a small difference between the null and the alternative diet, we are not going to have very much power to detect the difference. So if the wandering flabrat comes from a population whose mean is 140, reading up to the curve from 140 on the horizontal, we have approximately 90% power of properly classifying such a flabrat.

In summary, the power increases as the two population means become further apart. Intuitively, if the diet that the wandering flabrat is on is going to make him or her that much heavier, then it is going to be that much easier to distinguish him or her from the flabrat who is on the null diet. So the power increases as the delta increases, if all the while we hold alpha constant.

So what is causing the loss in power? It is the overlap—the region where both populations have representation. Will we always have overlap? Are we stuck with the amount of overlap we have? We saw that to decrease the overlap we can pull the curves apart (increase delta, the distance between the means). Another way, if we maintain the same distance, is to make the curves tighter around their means. This can be achieved by decreasing the standard deviations. Since we are stuck with the standard deviations because they are fixed in the population, how else can we decrease them?

This is reminiscent of the Central Limit Theorem. There we saw a decrease in the standard deviations—more precisely the standard errors—when we considered the distribution of the sample means and we allowed the sample size to increase. So extending this idea here, we can look at two distributions with means delta apart and decrease the overlap if instead of looking at the distribution of an individual; we looked at the distribution of means of individuals from both distributions.



Suppose instead of measuring just one flabrat we measure 2 and take their average.

What is the distribution of their mean?

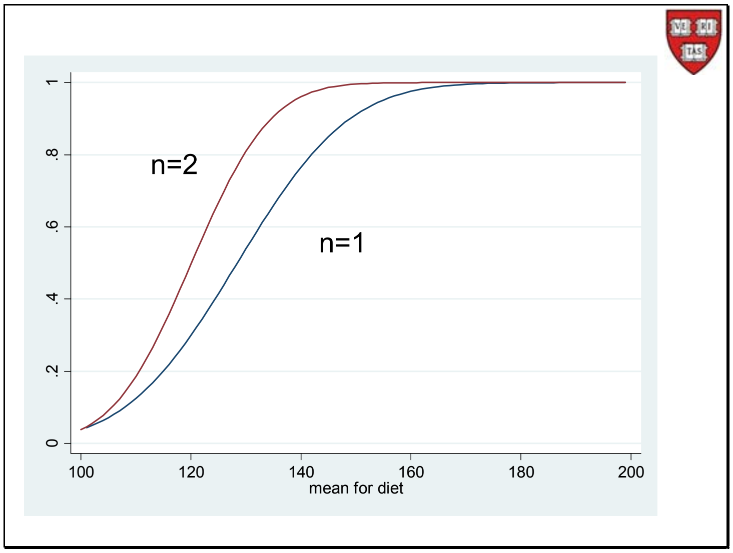
We know from the central limit theorem that the mean will be the same and the standard deviation will be reduced by a factor of $\sqrt{2}$

So in terms of our flabrats, if instead of finding one wandering flabrat out in the corridor we found 2 flabrats out wandering, then instead of weighing and classifying them individually, just like we did before for the single wanderer, assume that they both came from the same lab[3]. Then we can take their average, and classify them according to their average weight. (Remember from the Rice simulation that we saw much more variability in the top panel, the population, than we say in the third panel down, where we saw the distribution of the sample

[3] This is my story and I can make up the assumptions!

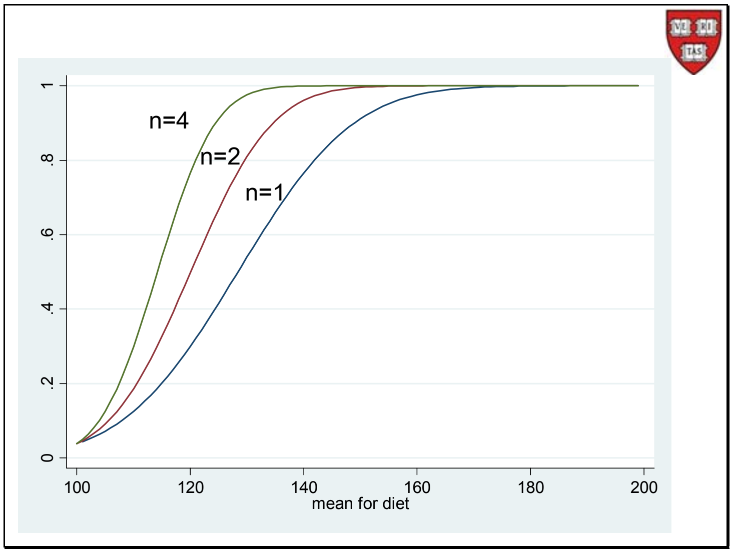
means—in the top panel the appropriate quantity to measure variability was the standard error, whereas in the third panel the appropriate quantity was the smaller standard error.)

Power and Sample Size



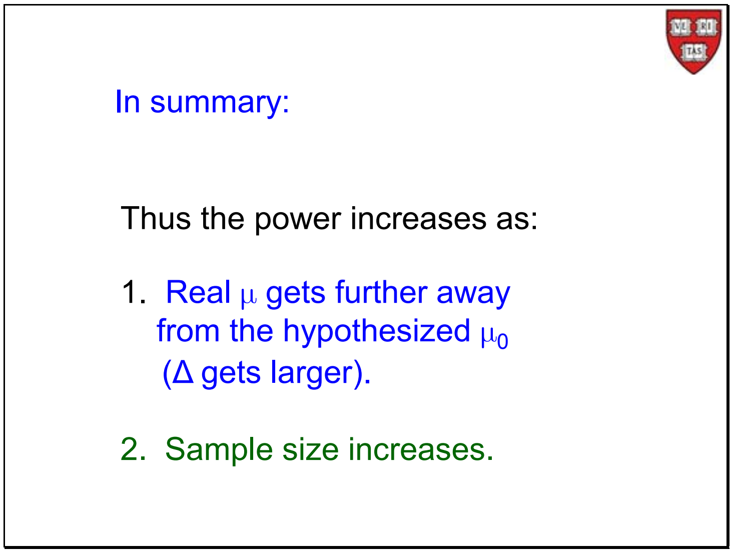
So now that you have convinced yourself that you can increase the power and lower the alpha, or keep the alpha constant, by increasing the sample size, you can see that the three are interrelated—namely, alpha, power (beta) and the sample size. Indeed, fixing two, fixes the third. (For those who like to think in these terms, that means we have three variables but only two degrees of freedom.)

For the moment, let us keep alpha constant, and look at the power curves and compare the power curve when the sample size is 1, to the power curve when the sample size is 2. We have precisely these two power curves plotted above. We can see, for example that when the mean of the alternative is 120 (which would be for the “Plus 20” diet) then we do not have a very high chance of properly classifying a wandering flabrat on this diet; indeed, the power is about 30%, or so. On the other hand, if 2 flabrats had escaped, then the power zooms up and almost doubles to about 60%.



If we had 4 wandering flabrats, so now we have a total prison outbreak, then let us superimpose the power curve for n=4. This curve dominates the other two curves to reflect an increase in power at every point of the domain, except, of course at the null, namely 100, because we have kept that at 5%.

So now even at 120 (the "Plus 20" diet), with 1 wandering flabrat we said the power it was about 30, but we go up to 4 wandering flabrats, then we have almost 90% power—let us say 87%.



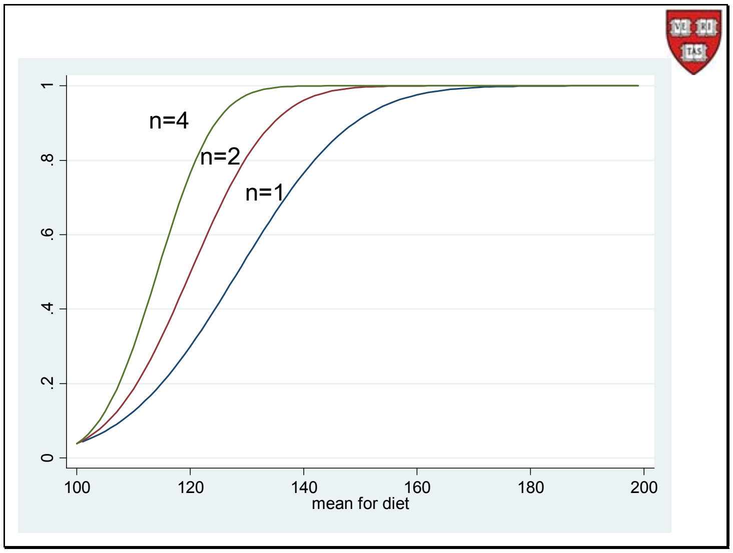
In summary:

Thus the power increases as:

1. Real μ gets further away from the hypothesized μ_0 (Δ gets larger).

2. Sample size increases.

So, in summary, from before we have that the power increases the more the two means of the distributions separate, and now we can add that the power also increases if the sample size increases.



Before leaving this graph let us make one more observation. We just read the graph by finding a point on the horizontal axis (such as 120) and shooting up and reading the values on the three curves. Now let us look at what happens if we look at first identifying a point on the vertical axis and looking horizontally until we hit the three curves.

For example, let's say we zero in on 0.8. We can ask, if I design my study to have 80% power, how big a difference between the means will I be able to detect? Of course, when we say “be able to detect” that has to be interpreted as detect with a certain power of detection. (No certainty in this course!) had power of 0.8, how small a difference would I be able to

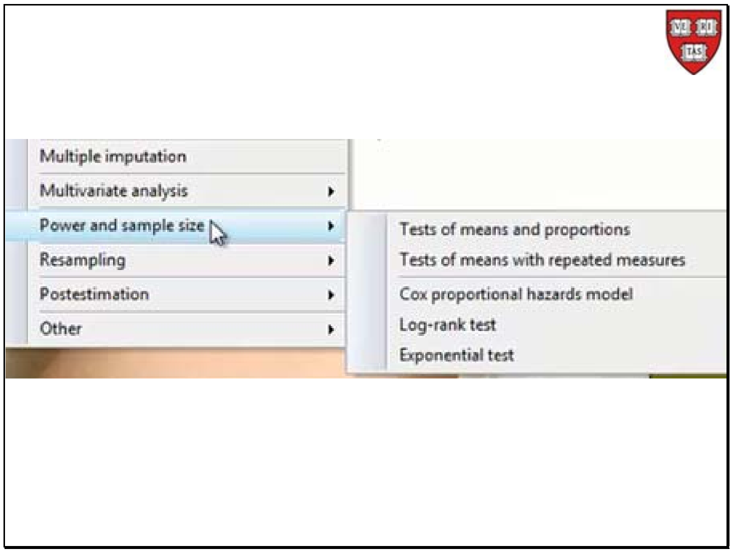
The answer is in the graph if we draw a horizontal line at 0.8. That line crosses the “n=4” curve at let us say 117; the “n=2” curve at 128; and, the “n=1” curve at 141. What that means is that we have 80% power to detect a “Plus 17” diet with a sample of size 4—actually we should say 17, or higher, because this power curve goes above 80% to the right of 117.

On the other hand, if we only have a sample of size 2, then we need to have a bigger separation of the means, namely to 128, or a “Plus 28” diet, to have an 80% chance of detecting a difference. That is, means of 128, or higher, to have at least an 80% chance of detecting a difference.

Finally, with a single wandering flabrat, to have at least an 80% chance of detecting the difference we need to have the alternative mean be 141, or higher.

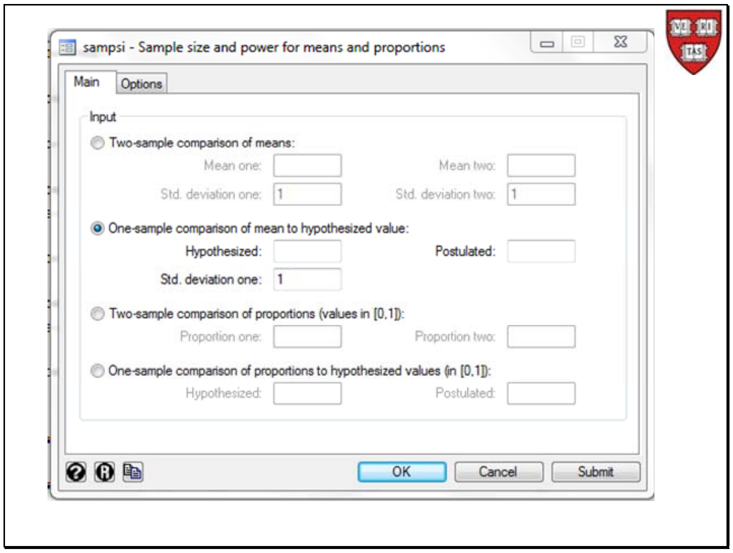
This is very often the way we design studies: namely, decide how big a difference you wish to detect—for example, what difference would be clinically meaningful—decide what power would be acceptable, and determine how big a sample size you will need to satisfy those constraints.

These considerations should remind us of the thinking we did when we looked at diagnostic testing. We had two kinds of errors and it was the accuracy of our measuring instrument that determined how precise our measurements could be. Here we can think of our measuring instrument as being the sample and the statistical analysis we do on that sample, and one way of buying a more precise instrument is to get a larger sample.



Let us turn to Stata to help us do the necessary calculations that guide us when designing a study. We first look at the one-sample situation and then at the two-sample case:

We start by clicking on "Statistics" and choosing "Power and sample size" and then "Tests of means and proportions", to get:



sampsi - Sample size and power for means and proportions

Main | Options

Input

○ Two-sample comparison of means:
Mean one: Mean two:
Std. deviation one: 1 Std. deviation two: 1

● One-sample comparison of mean to hypothesized value:
Hypothesized: Postulated:
Std. deviation one: 1

○ Two-sample comparison of proportions (values in [0,1]):
Proportion one: Proportion two:

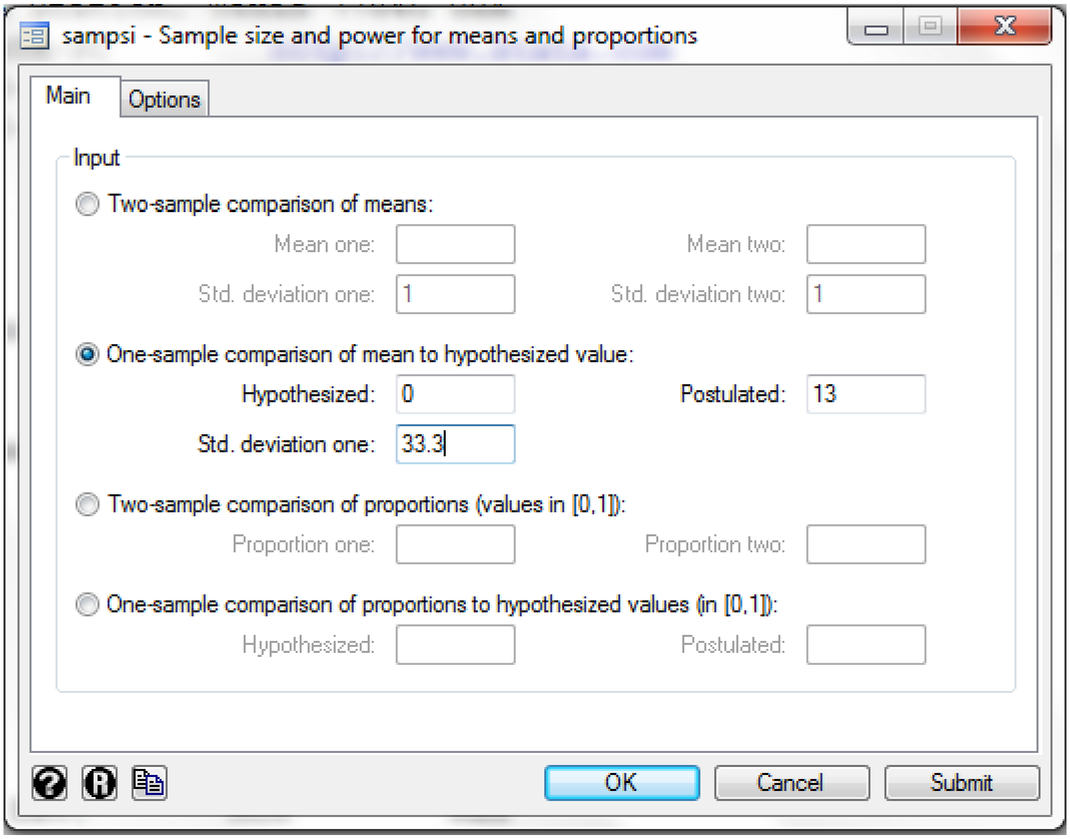
○ One-sample comparison of proportions to hypothesized values (in [0,1]):
Hypothesized: Postulated:

Let's look at the second one, "One-sample comparison of mean to hypothesized value:". Now what it wants from us are three quantities: the hypothesized value of the mean, the postulated value of the mean (where you want to calculate the power), and the standard deviation.

Before answering these questions let us set this aside as we set up an example.

```
. gen totchol = totchol2 - totchol1
(675 missing values generated)

. summ totchol

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     totchol |      3759    13.16574    33.30605       -159        321
```

We had previously defined *totchol* to be the difference between the total cholesterol at visit 2 minus the total cholesterol at visit 1 in the Framingham Heart Study. When we summarize this variable we find that its mean is 13 and its standard deviation is 33.



sampsi - Sample size and power for means and proportions

Input

○ Two-sample comparison of means:

Mean one: ____ Mean two: ____

Std. deviation one: 1 Std. deviation two: 1

● One-sample comparison of mean to hypothesized value:

Hypothesized: 0 Postulated: 13

Std. deviation one: 33.3

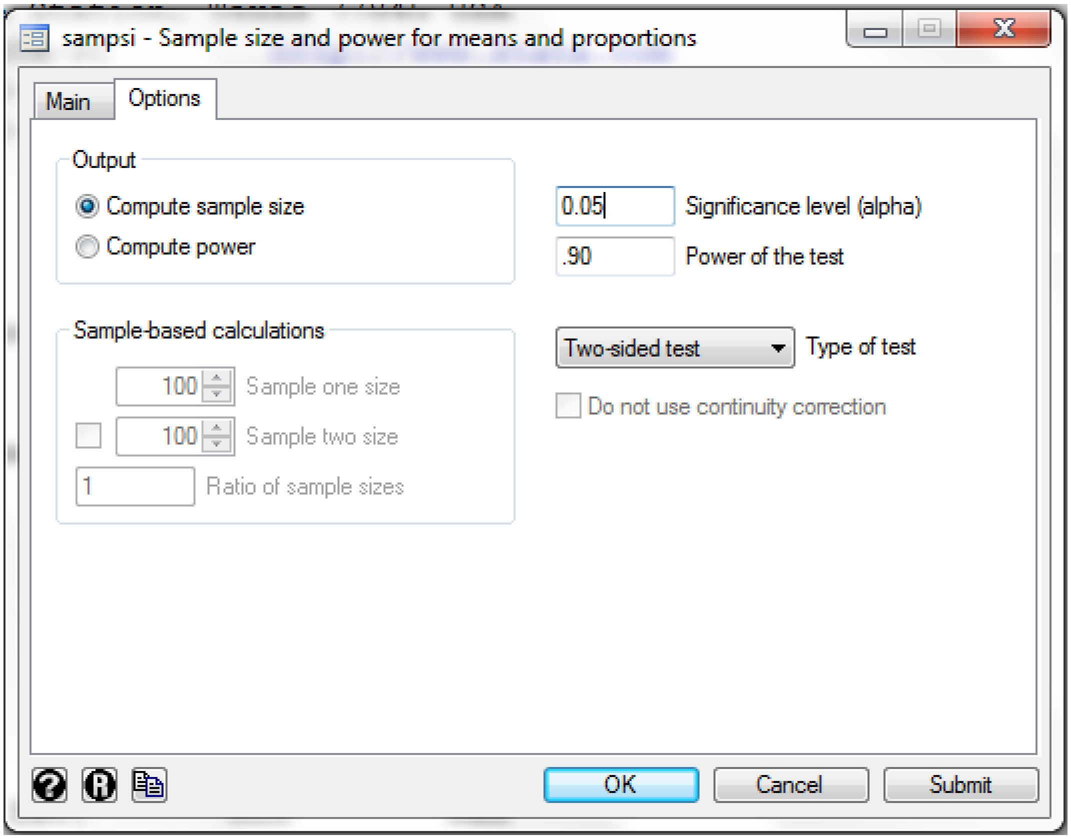
○ Two-sample comparison of proportions (values in [0,1]):

Proportion one: ____ Proportion two: ____

○ One-sample comparison of proportions to hypothesized values (in [0,1]):

Hypothesized: ____ Postulated: ____

Returning to our program, *sampsi*, let us fill in the hypothesized value as zero—so we are hypothesizing that there is no change in total cholesterol between the two visits—that the standard deviation is 33.3—ordinarily, here we need something that approximates the truth and one relies on past experiences with such observations, short of that, one cannot proceed—and then the postulated value. We have used 13, but this is the point where the power will be calculated. It should make subject-matter sense.



sampsi - Sample size and power for means and proportions

Output

● Compute sample size

○ Compute power

0.05 Significance level (alpha)

.90 Power of the test

Sample-based calculations

100 Sample one size

□ 100 Sample two size

1 Ratio of sample sizes

Two-sided test Type of test

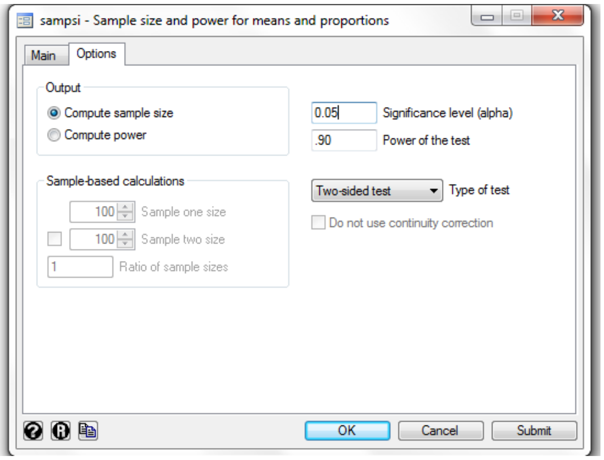
□ Do not use continuity correction

Now click on the “Options” tab to get this menu.

What alpha do we want? Let us put our friend 0.05 in for alpha, and it suggests a power of 90%, so let us leave that. And let us look at a two-sided test. And so what we're going to ask Stata to do for us is the compute the sample size.

Now when we click “Submit” we are asking the question, how big a sample do I need if I am going to test the hypothesis that the population mean is zero, at the alpha of 0.05 and I want a a 90% chance of detecting a difference of 13.

```
. sampsi 0 13, sd1(33.3) alpha(0.05) onesample

Estimated sample size for one-sample comparison of mean
  to hypothesized value

Test Ho: m =       0, where m is the mean in the population

Assumptions:

         alpha =   0.0500  (two-sided)
         power =   0.9000
 alternative m =       13
            sd =     33.3

Estimated required sample size:

             n =       69

.
```



sampsi - Sample size and power for means and proportions

Output

● Compute sample size

○ Compute power

0.05 Significance level (alpha)

.90 Power of the test

Sample-based calculations

100 Sample one size

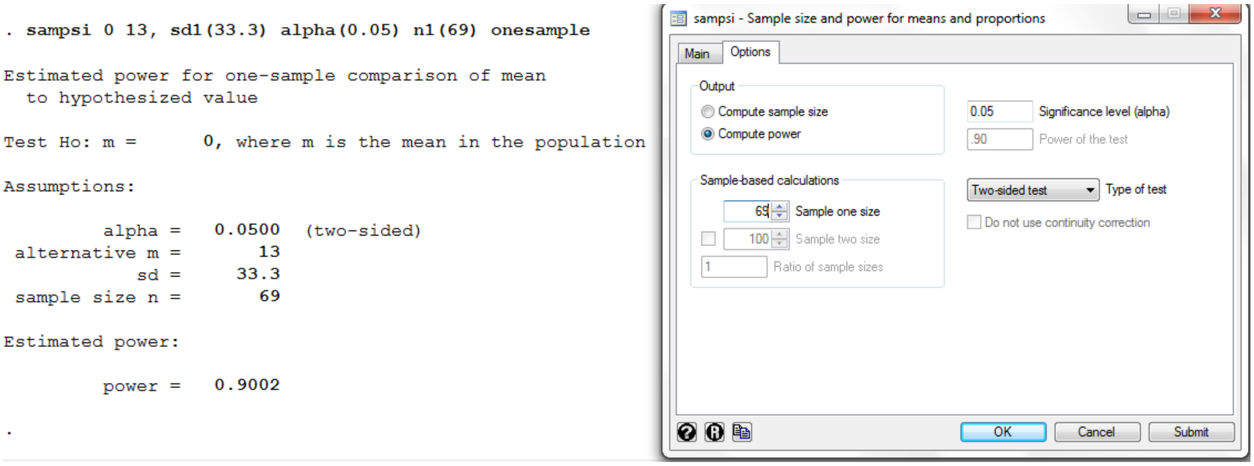
□ 100 Sample two size

1 Ratio of sample sizes

Two-sided test Type of test

□ Do not use continuity correction

Stata returns the value 69.



```
. sampsi 0 13, sd1(33.3) alpha(0.05) n1(69) onesample

Estimated power for one-sample comparison of mean
  to hypothesized value

Test Ho: m =       0, where m is the mean in the population

Assumptions:

         alpha =   0.0500  (two-sided)
 alternative m =       13
            sd =     33.3
 sample size n =       69

Estimated power:

         power =   0.9002

.
```

sampsi - Sample size and power for means and proportions

Output
○ Compute sample size
● Compute power

0.05 Significance level (alpha)
.90 Power of the test

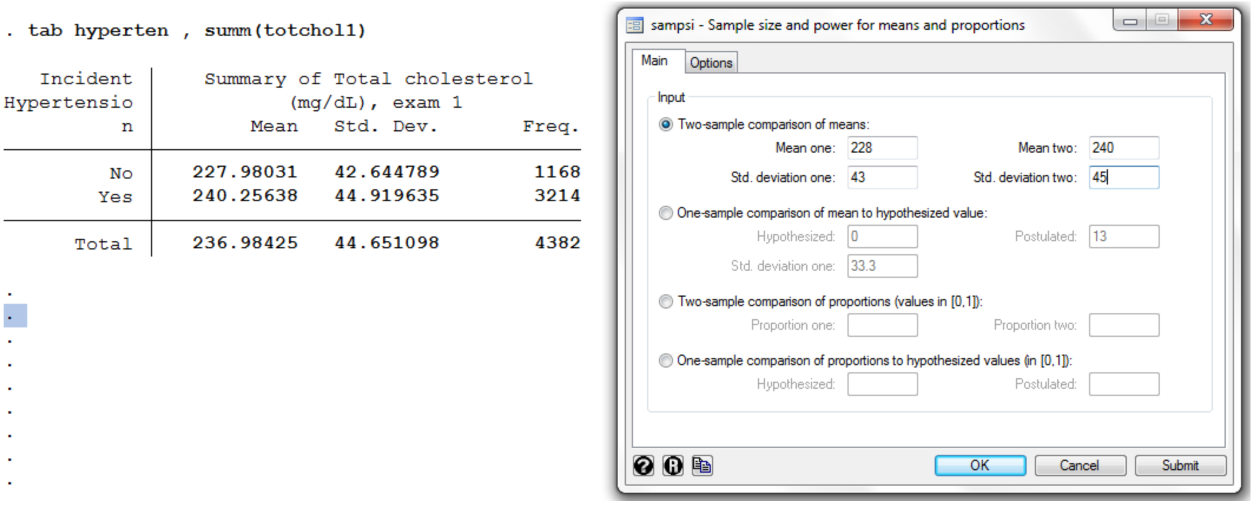
Sample-based calculations
69 Sample one size
□ 100 Sample two size
1 Ratio of sample sizes

Two-sided test Type of test
□ Do not use continuity correction

We could have done it the other way around asking Stata to compute the power for us (all the while at 13) when we have a sample of size 69 by clicking on “Compute power” and filling in the appropriate window in “Sample-based calculations”.

The answer we get—as expected—is 90.02%.

You can explore any number of “what if” scenarios by changing any of the numbers the program asks for, and get an idea of how uncertainty and precision vary together with the amount of knowledge needed etcetera.



```
. tab hyperten , summ(totchol1)

   Incident |     Summary of Total cholesterol
 Hypertensio |           (mg/dL), exam 1
           n |       Mean   Std. Dev.       Freq.
 ------------+------------------------------------
          No |  227.98031   42.644789        1168
         Yes |  240.25638   44.919635        3214
 ------------+------------------------------------
       Total |  236.98425   44.651098        4382

.
.
.
.
.
.
.
.
.
```

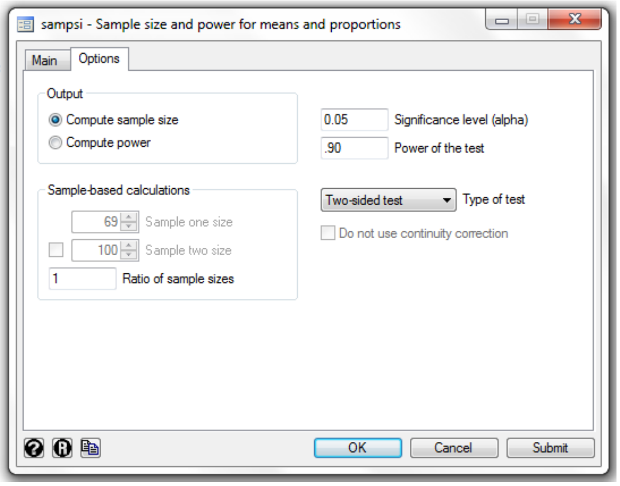
sampsi - Sample size and power for means and proportions

Input
● Two-sample comparison of means:
Mean one: 228 Mean two: 240
Std. deviation one: 43 Std. deviation two: 45
○ One-sample comparison of mean to hypothesized value:
Hypothesized: 0 Postulated: 13
Std. deviation one: 33.3
○ Two-sample comparison of proportions (values in [0,1]):
Proportion one: Proportion two:
○ One-sample comparison of proportions to hypothesized values (in [0,1]):
Hypothesized: Postulated:

Let us now look at the two sample comparison of means. We first looked at this when we looked at the difference in total cholesterol level at visit 1 between those who had had a hypertensive event and those who had not. Here is the summary of those two groups.

The means are 228 and 240, so let us use those means in Stata. The standard deviations for the two groups are at 43 and 45, so let us use those too.

```
.
. sampsi 228 240, sd1(43) sd2(45) alpha(0.05)

Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                    and m2 is the mean in population 2
Assumptions:

         alpha =   0.0500  (two-sided)
         power =   0.9000
            m1 =      228
            m2 =      240
           sd1 =       43
           sd2 =       45
         n2/n1 =     1.00

Estimated required sample sizes:

            n1 =      283
            n2 =      283
```



sampsi - Sample size and power for means and proportions

Main | Options

Output
- ● Compute sample size
- ○ Compute power

0.05 Significance level (alpha)

.90 Power of the test

Sample-based calculations
- 69 Sample one size
- □ 100 Sample two size
- 1 Ratio of sample sizes

Two-sided test Type of test

□ Do not use continuity correction

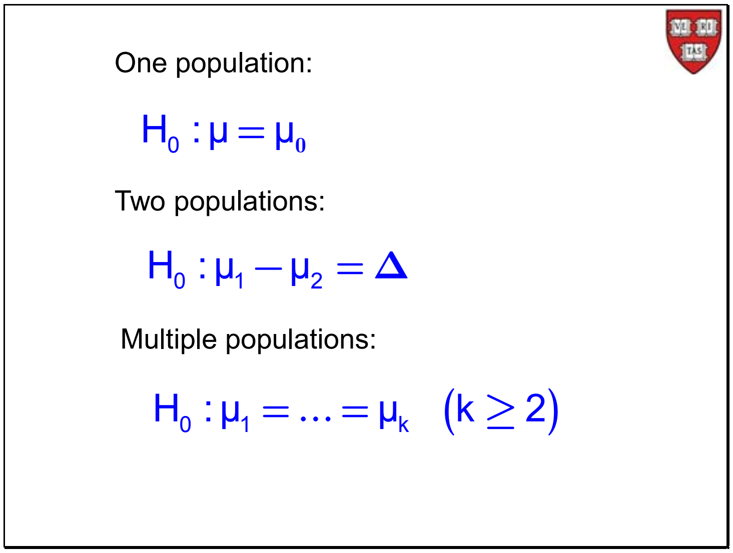
Now, let us look at our options. Let us leave the significance level at 0.05 and the power at 0.90 and leave the “Type of test” at Two-sided test, and submit that.

What we get back is that we are going to need a sample of size 283 from each of the two groups. (We can also play with the ratio of the two sample sizes, but I leave that to you to discover why you may want to do that.)

What if we want to spot a bigger difference, say between 228 and 250, then we go back to the previous menu and change the 240 to 250 and resubmit to get that we would need a much smaller sample of 85 from each of the groups.

These sample size calculations are very important when one is designing a study because they often dictate the difference between what is, and what is not a feasible study. Once again, I would recommend you explore different scenarios with this program to see how all these parameters are interrelated.

ANOVA



One population:

$$H_0 : \mu = \mu_0$$

Two populations:

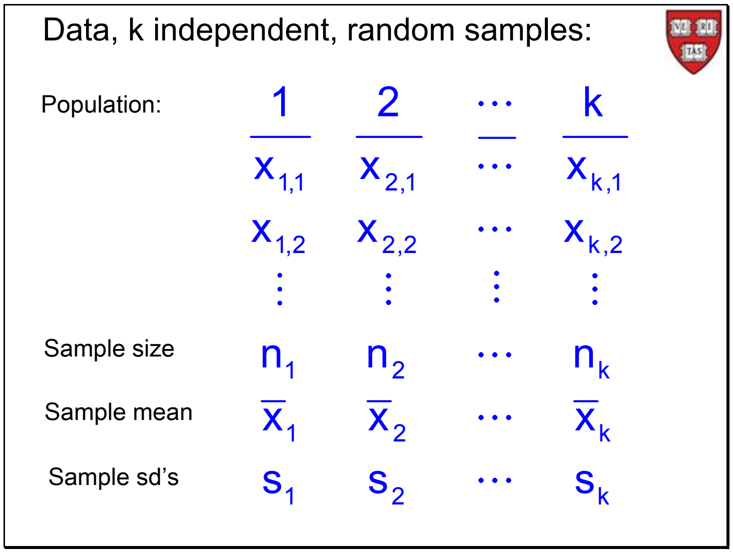
$$H_0 : \mu_1 - \mu_2 = \Delta$$

Multiple populations:

$$H_0 : \mu_1 = \ldots = \mu_k \quad (k \geq 2)$$

So far we have studied testing the mean of single population and then the means of two populations. What happens when we have more than two populations? This question was very useful in agricultural experimentations in the past when one was restricted to a single growing season per year, so it was beneficial to perform more than one experiment a year. In medicine we are faced with a similar problem when investigating bleak situations such as lung cancer where curative treatments are not forthcoming and one has to investigate a large number of potential treatments and there is some savings in time and effort in doing them simultaneously.

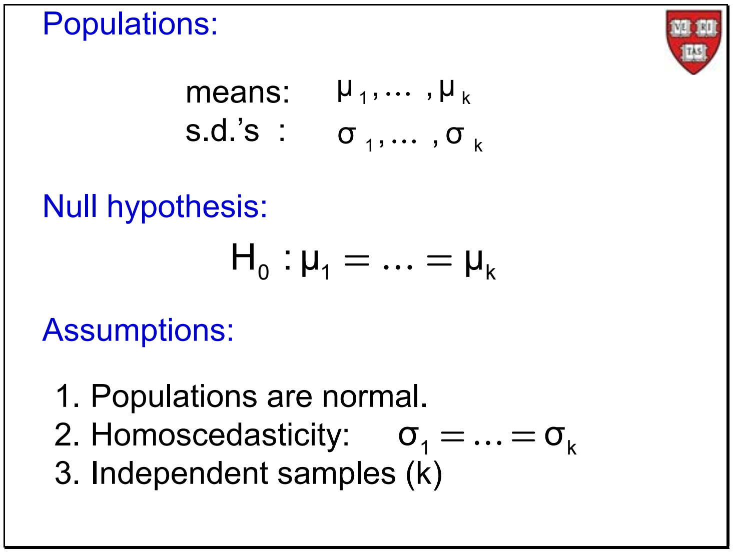
There are a number of ways to approach this problem, but I first want to concentrate on the simple one where we want to test the single hypothesis that the means of each of the populations are equal to each other.



Data, k independent, random samples:

Population:	1	2	$\cdots$	k
	$x_{1,1}$	$x_{2,1}$	$\cdots$	$x_{k,1}$
	$x_{1,2}$	$x_{2,2}$	$\cdots$	$x_{k,2}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Sample size	n_1	n_2	$\cdots$	n_k
Sample mean	$\bar{x}_1$	$\bar{x}_2$	$\cdots$	$\bar{x}_k$
Sample sd's	s_1	s_2	$\cdots$	s_k

So we start with k independent random samples. For each we have a sample size, a sample mean and a sample standard deviation.



Populations:

means: $\mu_1, \dots, \mu_k$

s.d.'s : $\sigma_1, \dots, \sigma_k$

Null hypothesis:

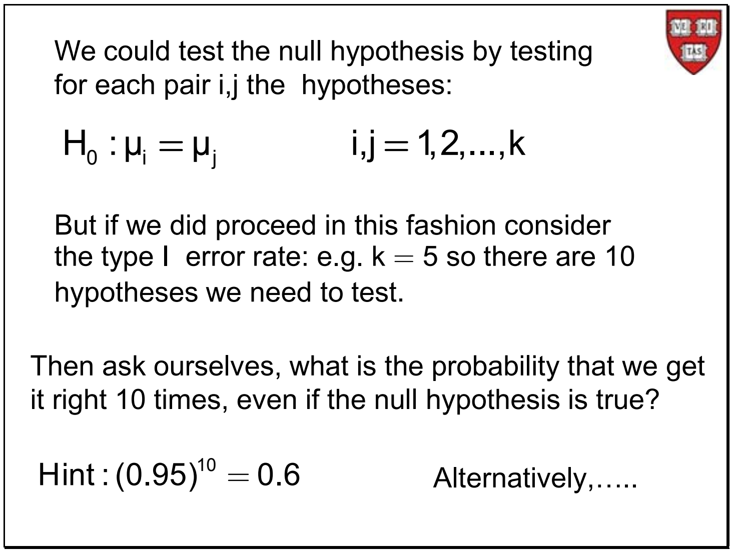
$$H_0 : \mu_1 = \dots = \mu_k$$

Assumptions:

1. Populations are normal.
2. Homoscedasticity: $\sigma_1 = \dots = \sigma_k$
3. Independent samples (k)

The k populations have means and standard deviations, and we wish to test the hypothesis that these k population means are all equal to each other.

We only look at the situation when all k populations are normally distributed, the populations are homoscedastic, and the samples are independent. One could remove these assumptions, but that is beyond the scope of this course.



We could test the null hypothesis by testing for each pair i,j the hypotheses:

$$H_0 : \mu_i = \mu_j \qquad i,j = 1,2,\dots,k$$

But if we did proceed in this fashion consider the type I error rate: e.g. k = 5 so there are 10 hypotheses we need to test.

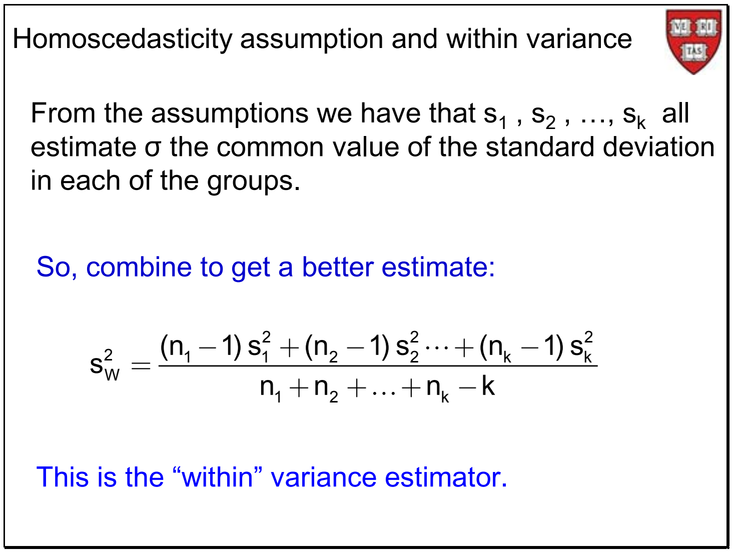
Then ask ourselves, what is the probability that we get it right 10 times, even if the null hypothesis is true?

$\text{Hint}: (0.95)^{10} = 0.6$ Alternatively,…..

One way to test the hypothesis that all k population means are equal is to use what we have developed so far in the course and compare all the populations pairwise; since we know how to test the equality of the means of two populations.

Think a little about what would happen to your error rate if you proceed in this fashion. For example if k is equal to 5, then you would have 5 combination 2, or 10 tests to do. What is the probability that at least one of these is wrong. That is one minus the probability that you get all 10 right. If all ten tests were independent and each was tested at alpha is 0.05, then the probability that you get at least one wrong, if the null hypothesis is true, is $(1\text{-}0.95^{10}) = 0.40$. Thus your 5% has ballooned to 40%. Granted not all the tests are going to be independent, but in some sense then things might be even worse.

So let us hope that this is not the best approach we can take. There is another approach called the *analysis of variance*, and as the name implies, we study the behavior of the sample variances to direct us to an answer.



Homoscedasticity assumption and within variance

From the assumptions we have that s_1 , s_2 , …, s_k all estimate σ the common value of the standard deviation in each of the groups.

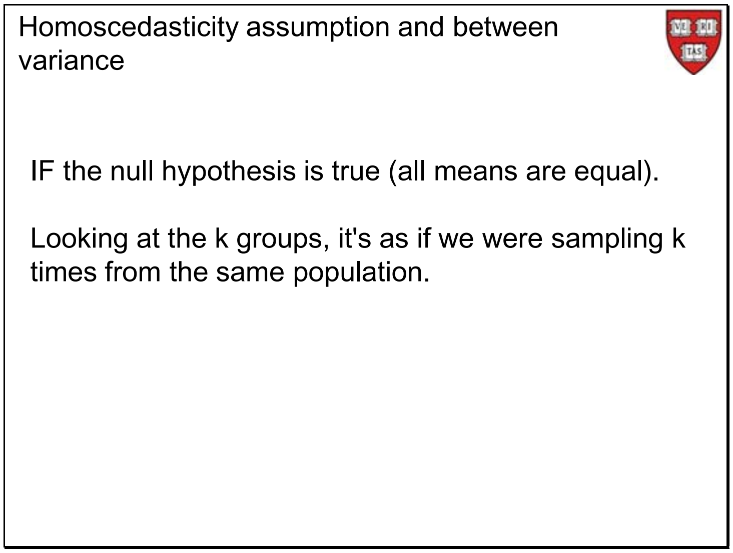
So, combine to get a better estimate:

$$s_W^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2 \cdots + (n_k - 1)\, s_k^2}{n_1 + n_2 + \ldots + n_k - k}$$

This is the “within” variance estimator.

Consider the assumption we have made of homoscedasticity. If that assumption is correct, then we have k estimates of a common quantity; that being the common population variance. Call it σ^2. We can estimate it by taking a weighted average (since each of the n_i may be different) of each of the sample variances. Call this estimate, s^2_w. This is called the “within” variance estimator (generated from within the k samples).

To justify this estimator we call on the homoscedasticity assumption. When we test a hypothesis we usually start with, if the null hypothesis is true we expect to see… and proceed from there.

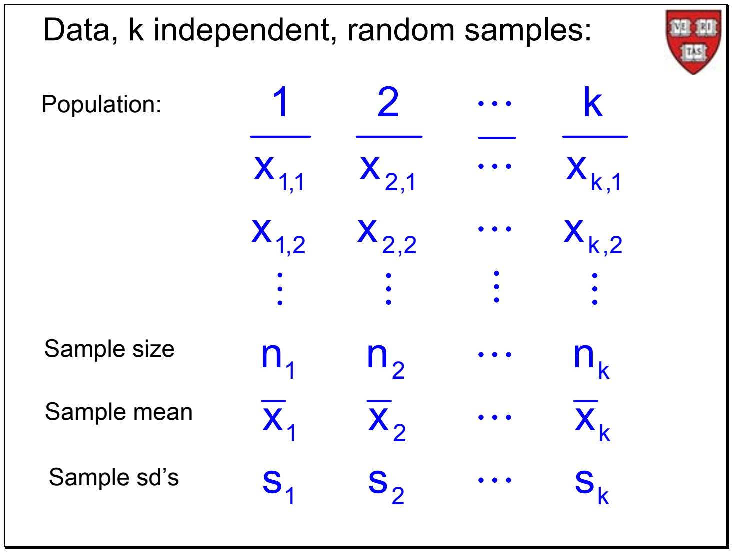


Homoscedasticity assumption and between variance

IF the null hypothesis is true (all means are equal).

Looking at the k groups, it's as if we were sampling k times from the same population.

So what do we expect to see if the null hypothesis is true? So if the null hypothesis is true, we expect all the means to be equal.

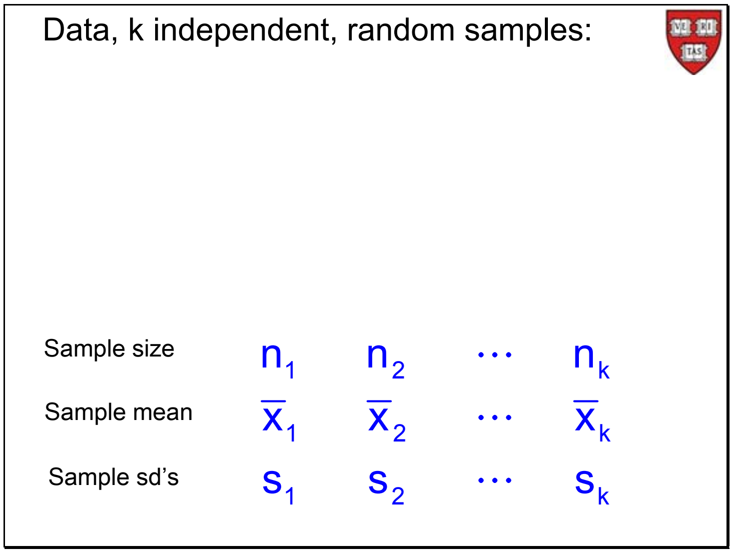


Data, k independent, random samples:

Population:	1	2	$\cdots$	k
	$x_{1,1}$	$x_{2,1}$	$\cdots$	$x_{k,1}$
	$x_{1,2}$	$x_{2,2}$	$\cdots$	$x_{k,2}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Sample size	n_1	n_2	$\cdots$	n_k
Sample mean	$\bar{x}_1$	$\bar{x}_2$	$\cdots$	$\bar{x}_k$
Sample sd's	s_1	s_2	$\cdots$	s_k

So let us look at these k groups. If the means are equal and we have homoscedasticity, then it's as if we were sampling k times from the same (normal) population.

The sample sizes may be different.

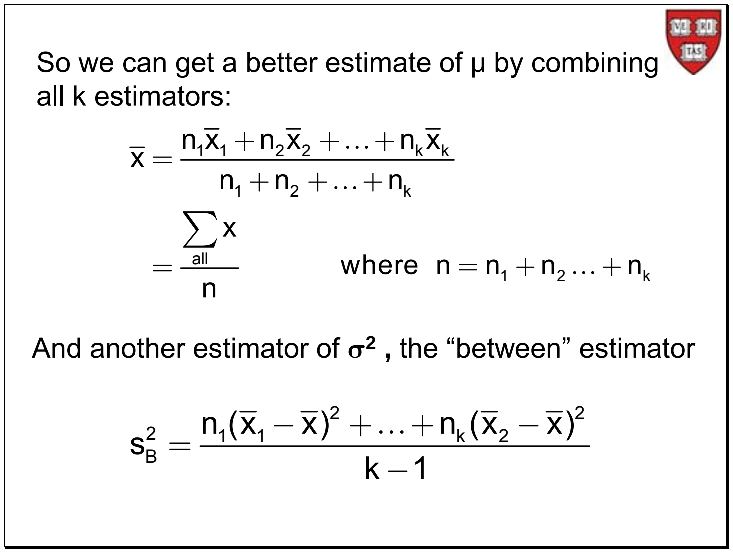


Data, k independent, random samples:

Sample size	n_1	n_2	$\cdots$	n_k
Sample mean	$\overline{x}_1$	$\overline{x}_2$	$\cdots$	$\overline{x}_k$
Sample sd's	s_1	s_2	$\cdots$	s_k

Let us focus on the summary statistics. Possibly different sample sizes, but each of these sample means is estimating the same overall mean (under the null) and each of these sample standard deviations is estimating the same standard deviation (homoscedasticity assumption).

So from the central limit theorem, $\overline{X}_1$ is a sample from a population that has mean μ and standard deviation $\sigma / \sqrt{n_1}$, $\overline{X}_2$ a sample from a population that has mean μ and standard deviation $\sigma / \sqrt{n_2}$, and so on. Both of the first two will give me information about σ . Indeed, all k of them will. Combining these k we can construct a combined estimate of σ^2



So we can get a better estimate of μ by combining all k estimators:

$$\overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + \ldots + n_k\overline{x}_k}{n_1 + n_2 + \ldots + n_k}$$

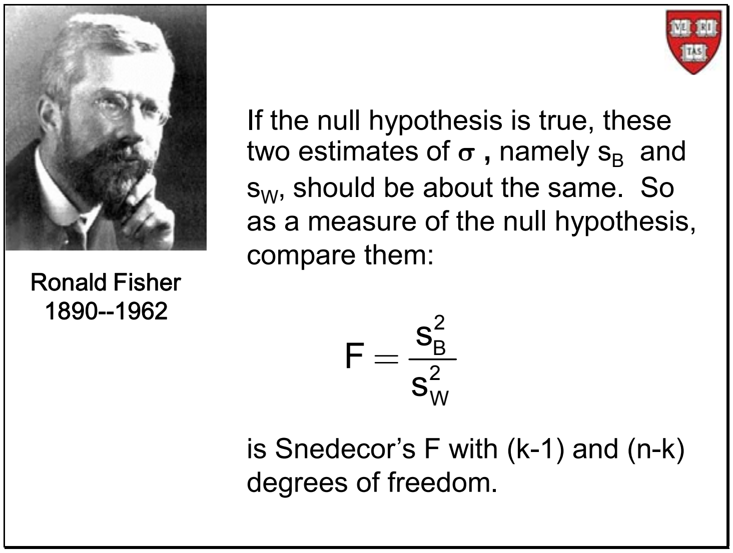
$$= \frac{\sum_{\text{all}} x}{n} \qquad \text{where } n = n_1 + n_2 \ldots + n_k$$

And another estimator of σ^2 , the "between" estimator

$$s_B^2 = \frac{n_1(\overline{x}_1 - \overline{x})^2 + \ldots + n_k(\overline{x}_2 - \overline{x})^2}{k-1}$$

First we combine them all to get an estimate of the common μ. We want to construct a weighted mean because we want to weight the larger sample sizes more than the small ones. In fact what we have done is tantamount to ignoring what sample the observations comes from and just summing them all up and dividing by how many observations we have; our usual way of calculating the mean.

We can now see how the individual $\bar{X}$ vary around the sample mean, and that should be related to the standard error. (Remember the definition of the standard error, it tells us how much the sample means vary around the overall mean.) We call this variance estimator the between estimator, s^2_B.



Ronald Fisher
1890--1962

If the null hypothesis is true, these two estimates of σ , namely s_B and s_W, should be about the same. So as a measure of the null hypothesis, compare them:

$$F = \frac{s_B^2}{s_W^2}$$

is Snedecor's F with (k-1) and (n-k) degrees of freedom.

The genius of it all is to now just compare these estimators of the same quantity. We can take their ratio. It is called the F statistic[4]. This should be approximately 1 if the null-hypothesis is true. So just like we have had statistics and their distributions, for example the Z and the t, we now have the F and we can use it to determine p-values we can attach to a hypothesis.

Let us take a look at an example of the analysis of variance (ANOVA).

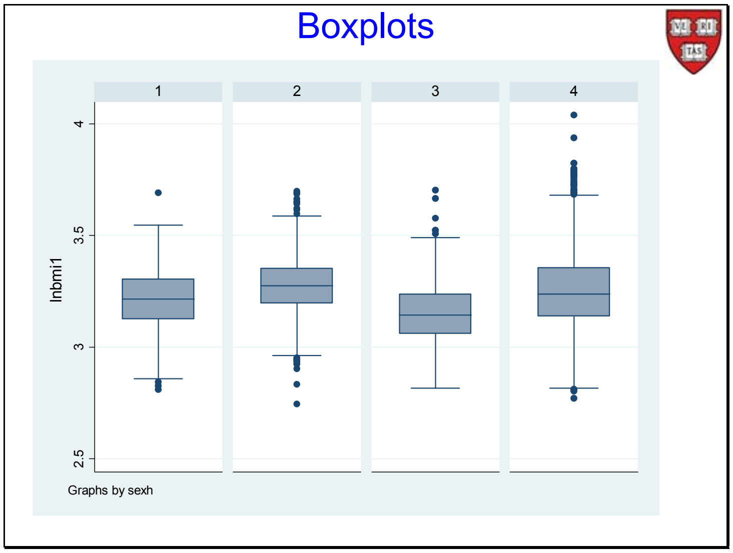
I have created a new variable, and called it *sexh*. And what this variable does is it breaks up our population into four groups. The first two refer to men, and the bottom two refer to women. And what the h stands for is hypertension. So I'm going to look at incident hypertension.

And so the first group (sexh=1) is going to be men without hypertension; the second (sexh=2) is men with hypertension; the third group (sexh=3) is women without hypertension; and the fourth

[4] The statistic is related to one Fisher had proposed in one of the most important papers about ANOVA he wrote in the 1920s, but this one was proposed and tabulated by George Snedecor who named it F in honor of Fisher. Apparently Fisher was none too pleased to have seen his statistic modified.

group (sexh=4) is women with hypertension. What we would like to find out is whether there is a difference between these four groups, with respect to an outcome.

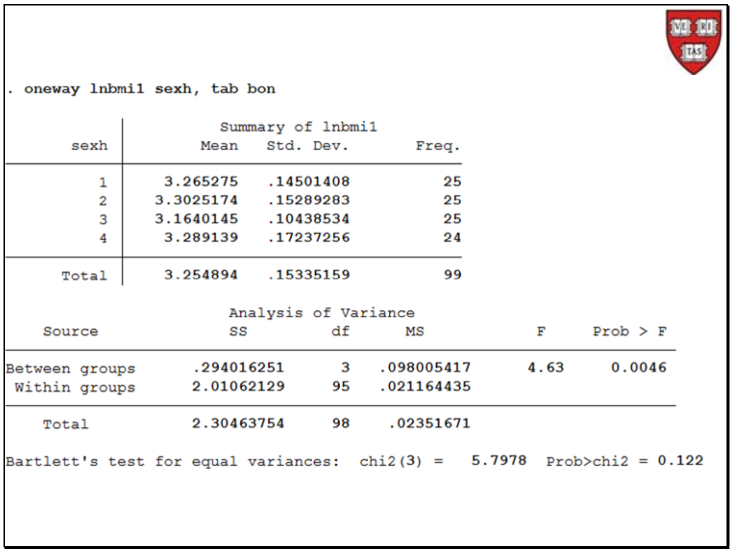
Let us define a new outcome, the logarithm of BMI. Remember, with the analysis of variance, we need to assume normality, so that is why we look at the logarithm of BMI; it is closer to being normally distributed than the raw BMI.



Boxplots

To get a feel for the data, let us look the box plots of these groups. From the boxplot, it looks like the BMI is slightly higher for men and women with hypertension, and it looks like the difference is bigger for women than it is for men.

Let's see what happens if we take a sample from this population, and submit it to the an ANOVA. So let us take a sample of size 25 from each of these four groups.



```
. oneway lnbmi1 sexh, tab bon

            |          Summary of lnbmi1
       sexh |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |    3.265275   .14501408          25
          2 |   3.3025174   .15289283          25
          3 |   3.1640145   .10438534          25
          4 |    3.289139   .17237256          24
------------+------------------------------------
      Total |    3.254894   .15335159          99

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      .294016251      3   .098005417      4.63     0.0046
 Within groups      2.01062129     95   .021164435
------------------------------------------------------------------------
    Total           2.30463754     98    .02351671

Bartlett's test for equal variances:  chi2(3) =   5.7978  Prob>chi2 = 0.122
```

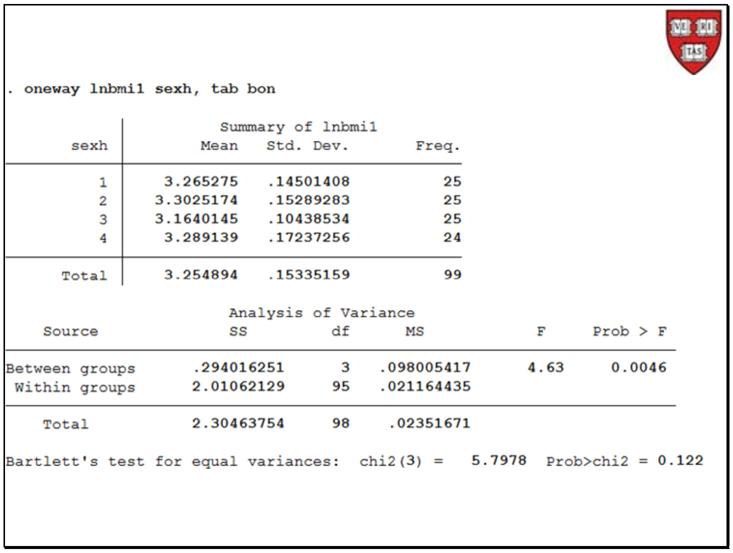
The Stata command for the analysis of variance is *oneway*—there exist more complex analyses, this is the simplest one, and that is why it is called oneway.

We see that we lost one person in group 4; a missing value. Let us ignore that. The means and standard deviations for the four groups are reported. We see, in agreement with what we saw with the boxplots for the whole population, that those without hypertension (groups 1 and 3) have lower BMIs (log BMIs to be precise) , and that women do better than men (have lower BMIs).

Here is the analysis of variance table. The column labeled MS gives us the two quantities we called s^2_B and s^2_W, above. Their ratio is in the column labeled F, and the p-value associated with the null hypothesis of equality of the four population means is given in the last column, 0.0046.

So at the 0.05 level we reject the null hypothesis of the equality of the population means.

The last line also gives us the results of Bartlett's test for homoscedasticity. This test is somewhat sensitive to departures from normality, but there it is and what it tells us is that the data seems to be consonant with a hypothesis that we have homoscedasticity. In other words our assumption seems to be acceptable.

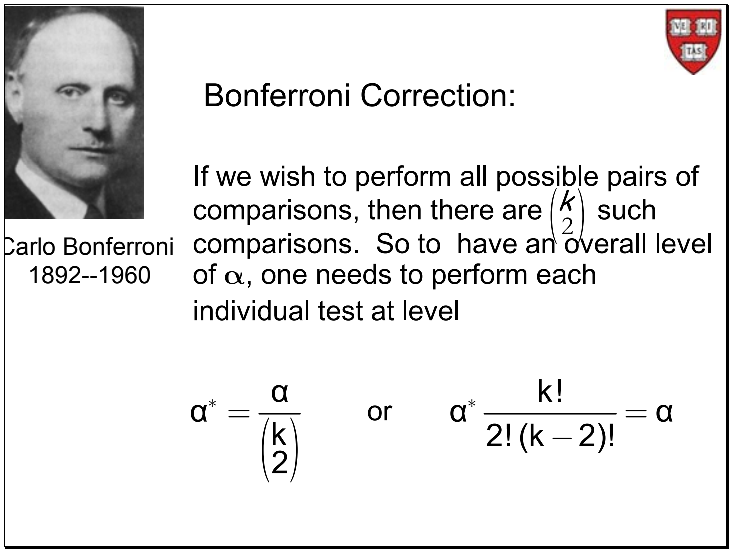


```
. oneway lnbmi1 sexh, tab bon

            |          Summary of lnbmi1
       sexh |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |    3.265275   .14501408          25
          2 |   3.3025174   .15289283          25
          3 |   3.1640145   .10438534          25
          4 |    3.289139   .17237256          24
------------+------------------------------------
      Total |    3.254894   .15335159          99

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      .294016251      3   .098005417      4.63     0.0046
 Within groups      2.01062129     95   .021164435
------------------------------------------------------------------------
    Total           2.30463754     98    .02351671

Bartlett's test for equal variances:  chi2(3) =   5.7978  Prob>chi2 = 0.122
```

So where we left it was at this point where we said we will reject the null hypothesis that all the means are equal. Now here are all the means we observed, and you can see that group one and four look very close together, and both are close to group two. So what has caused us to reject the overall hypothesis?

Now remember the overall hypothesis was that all the means are equal. So any departure from this overall equality could be the cause of us rejecting the whole. It would be interesting to find out the cause(s).

There is something we can do. As an option in the *oneway* command I added comma *bon*. What did that give us?



Carlo Bonferroni
1892--1960

Bonferroni Correction:

If we wish to perform all possible pairs of comparisons, then there are $\binom{k}{2}$ such comparisons. So to have an overall level of α, one needs to perform each individual test at level

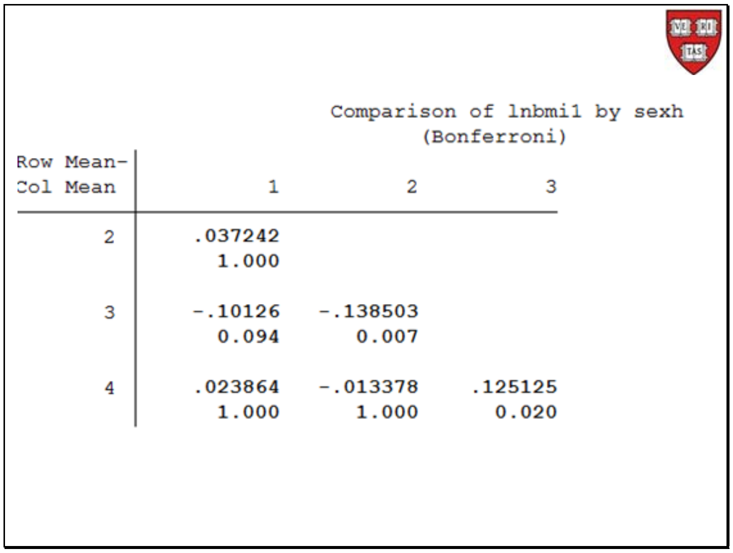
$$\alpha^* = \frac{\alpha}{\binom{k}{2}} \qquad \text{or} \qquad \alpha^* \frac{k!}{2!\,(k-2)!} = \alpha$$

The word bon is an abbreviation for the Bonferroni Correction. There are other ones we could use, but this one is quite conservative, and thus might be the best to use.

If we wish to perform all possible pairs of comparisons—so for example, 1 versus 2, and 1 versus 3, and 1 versus 4, and 2 versus 3, and 2 versus 4, and 3 versus 4, thus all the pairwise comparisons—then way to do that is to take your alpha, let's say 0.05, and divide it by k combination 2; the total number of possible pairwise comparisons.

So in this case k was 4, so k combination 2 is 6. So divide alpha by 6, and then test each one of these pairwise comparisons at this (alpha/6) level, to still maintain an overall level of alpha.

Or if you multiply by k combination 2, that is your overall alpha, and this is what Stata reports for us.



```
                    Comparison of lnbmi1 by sexh
                             (Bonferroni)
Row Mean-|
Col Mean |          1           2           3
---------+------------------------------------
       2 |    .037242
         |      1.000
         |
       3 |    -.10126   -.138503
         |      0.094      0.007
         |
       4 |    .023864   -.013378    .125125
         |      1.000      1.000      0.020
```

So what the call to *bon* did when we ran this command with Stata is to produce this table. In the table we look at the couplet in each row/column combination. The upper number is the mean of the group identified by the row label minus the mean of the group identified by the column label. The lower number is the modified p-value (to accommodate the Bonferroni correction) associated with the test of the hypothesis that the difference between the groups identified by the row/column identifiers, is zero.

So suppose we wish to maintain our overall alpha of 0.05, then the only two pairs for whom we would reject the null hypothesis of no difference in the means are the pairs, 2 and 3, and 3 and 4—because those two p-values are the only ones less than 0.05.

So the one group that is sticking out is group number 3, the women who are not hypertensive. They are not different from men who are not hypertensive, but they are significantly different from both the men and the women who are hypertensive. No pair of mean differences amongst the other three groups is found to be significantly different from zero. So the non-hypertensive women seem to be the cause of the rejection of the overall null hypothesis in the original analysis of variance.