UNIVERSITY OF CALGARY

Clinical Research Module: Introductory statistics

Lecture 1B

Gerry Giesbrecht, PhD, RPsych

These lecture notes based on others developed by Drs. Alberto Nettel-Aguirre, Peter Faris, Sarah Rose Luz Palacios-Derflingher and myself

UNIVERSITY OF CALGARY

Material Covered in this Lecture

- Lecture 1B- Inference
  - The logic and use of statistics
  - Sampling Distributions
    - The normal distribution
    - Confidence intervals
  - Estimation and Hypothesis testing
    - Difference of Means
  - Power and Sample Size

UNIVERSITY OF CALGARY

Lies, damn lies, and statistics

Statistics rely on assumptions and only provide approximate truths about the things that we really want to know about.

And that doesn't even get into the quality of your data, their representativeness…

UNIVERSITY OF CALGARY

So why bother?

Because the alternatives are worse.

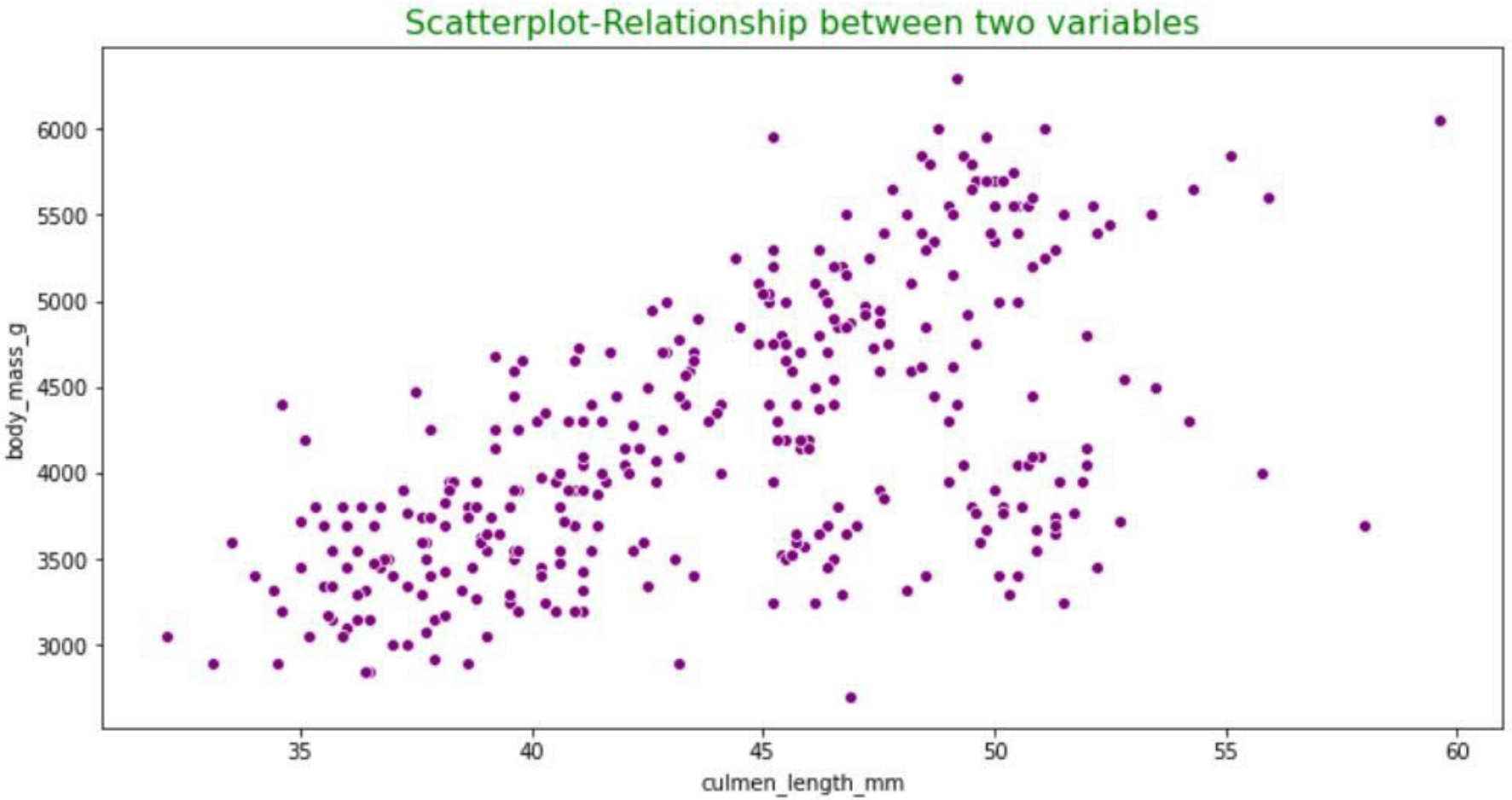
Because clinical wisdom/intuition is susceptible to chance and confirmation bias.

Because if we let the data speak, we will make new discoveries.

Because statistics help to separate the signal from the noise.

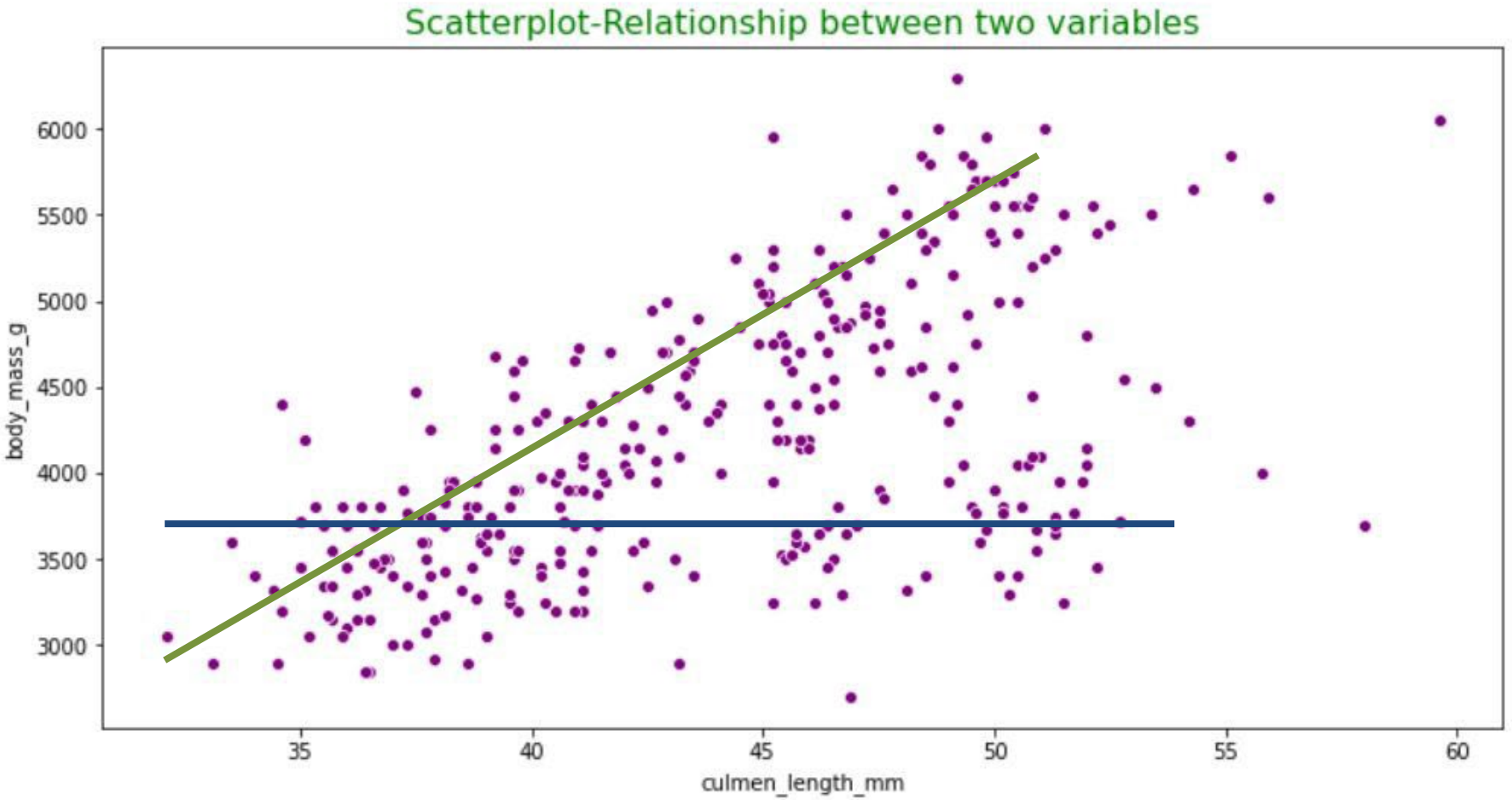
UNIVERSITY OF CALGARY

Example



UNIVERSITY OF CALGARY

Example



UNIVERSITY OF CALGARY

Statistics make knowledge feasible

Most of what we want to know, is (practically) unknowable.

E.g., what proportion of people who smoke for at least 2 years die of lung cancer?

Obtaining a sample of smokers and calculating the portion who die of lung cancer is feasible (but conclusions depend on the quality of the sample).

Example Results

Anesthesiology 2006; 105:665–9

that both the calculated and the standard total analgesic morphine doses in the < 85% group were lower ($P < 0.05$) than the calculated total analgesic morphine dose in the ≥ 85% saturation nadir group (fig. 2A). There was

Int J Pediatr Endocrinol 2011 Nov 8;2011(1):15.
…No significant differences were seen in the mean HbA1c between control and intervention at 0 months [8.48(0.86) vs 8.57(1.13)], F(1, 209) = 1.67, p = 0.43 ….

J Bone Joint Surg Am. 2012 Oct 17;94(20):1853-60.
…at six weeks, the children who were allowed to bear weight as tolerated had better overall scores (95% CI = 2.34, 4.33) and better standing skills (95% CI = 7.29, 8.89) than those who were initially instructed to be non-weight-bearing

Jacob Cohen: The earth is round p < .05.

UNIVERSITY OF CALGARY

The bottom line…

The two most important things to understand from the first part of this lecture:

- What is the best estimate for an effect (i.e., the point estimate)
- How much error is in the estimate (i.e., the confidence interval)

Material Covered in the Biostatistics Lectures

- Lecture 1B- Inference
  - The logic and use of statistics
  - Sampling Distributions
    - The normal distribution
    - Confidence intervals
  - Estimation and Hypothesis testing
    - Difference of Means
  - Power and Sample Size

UNIVERSITY OF CALGARY

Theoretical concepts

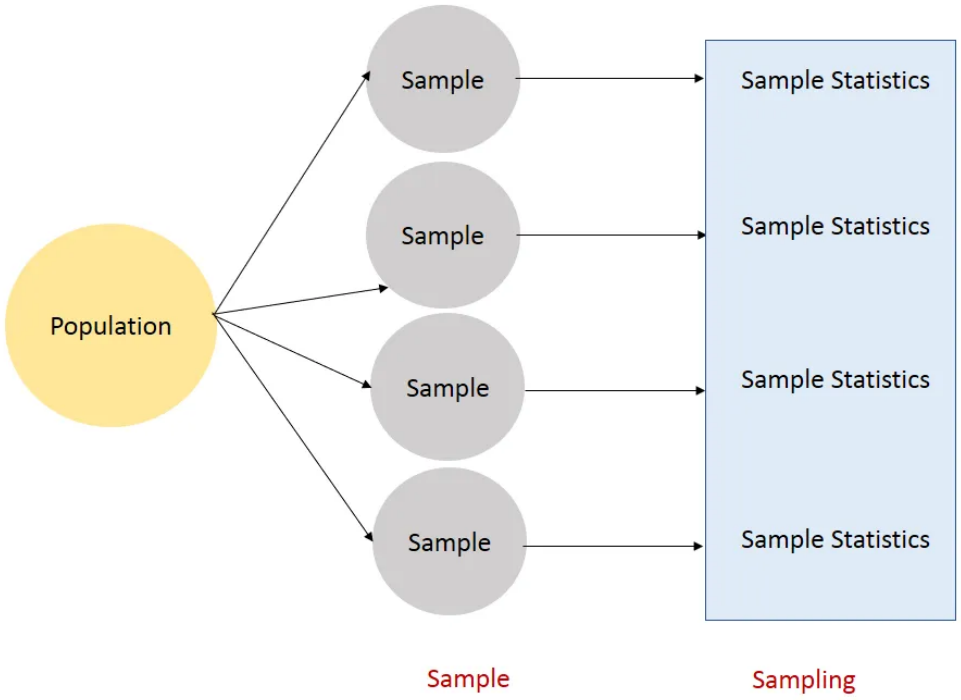
This might get a bit theoretical…

To understand statistics, we need to understand how the data we collect and the numbers we generate using statistics help us make sense of what is going on.

This requires some theory and theoretical concepts.

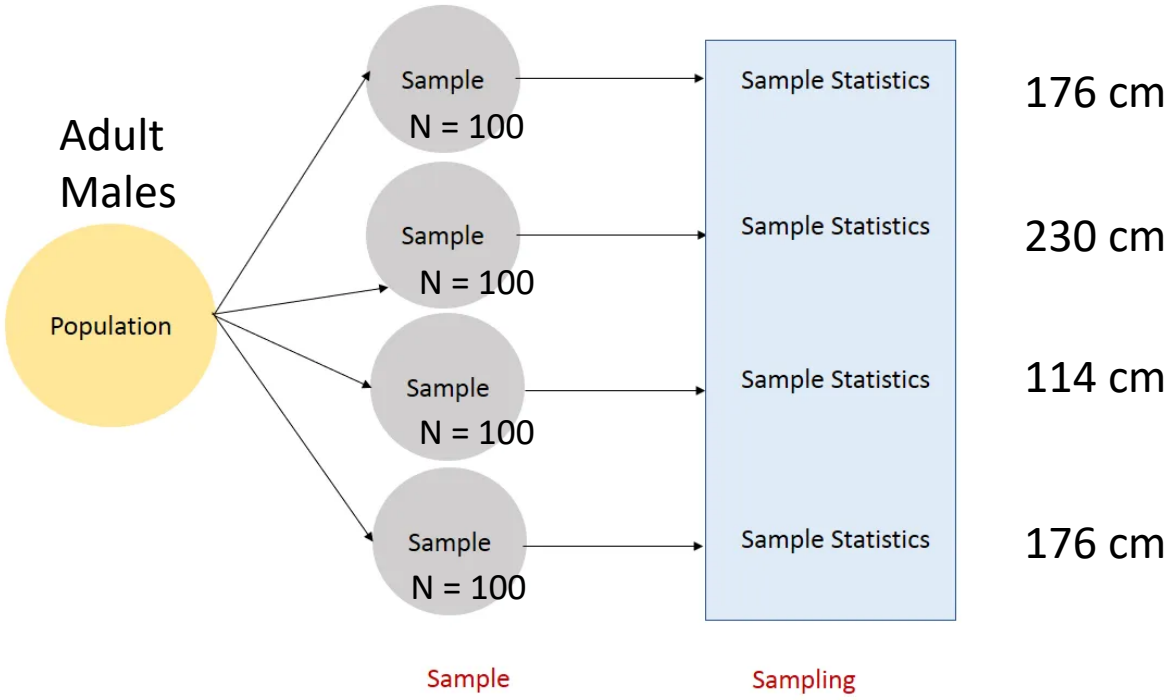
Sampling Distributions

- For any estimate, the **sampling distribution** is the distribution of that estimate calculated for all possible samples of size *n* drawn from a population.



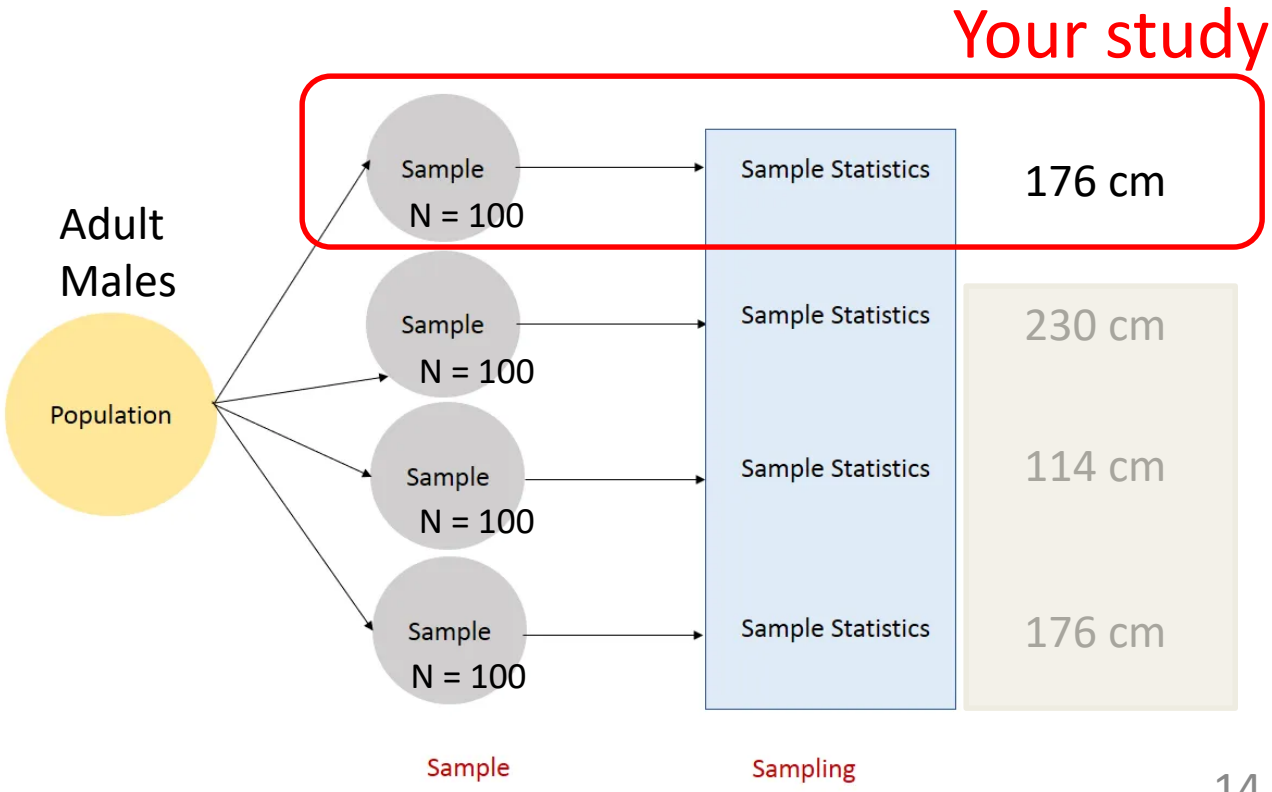
Sampling Distributions (example)

- For any estimate, the **sampling distribution** is the distribution of that estimate calculated for all possible samples of size *n* drawn from a population.



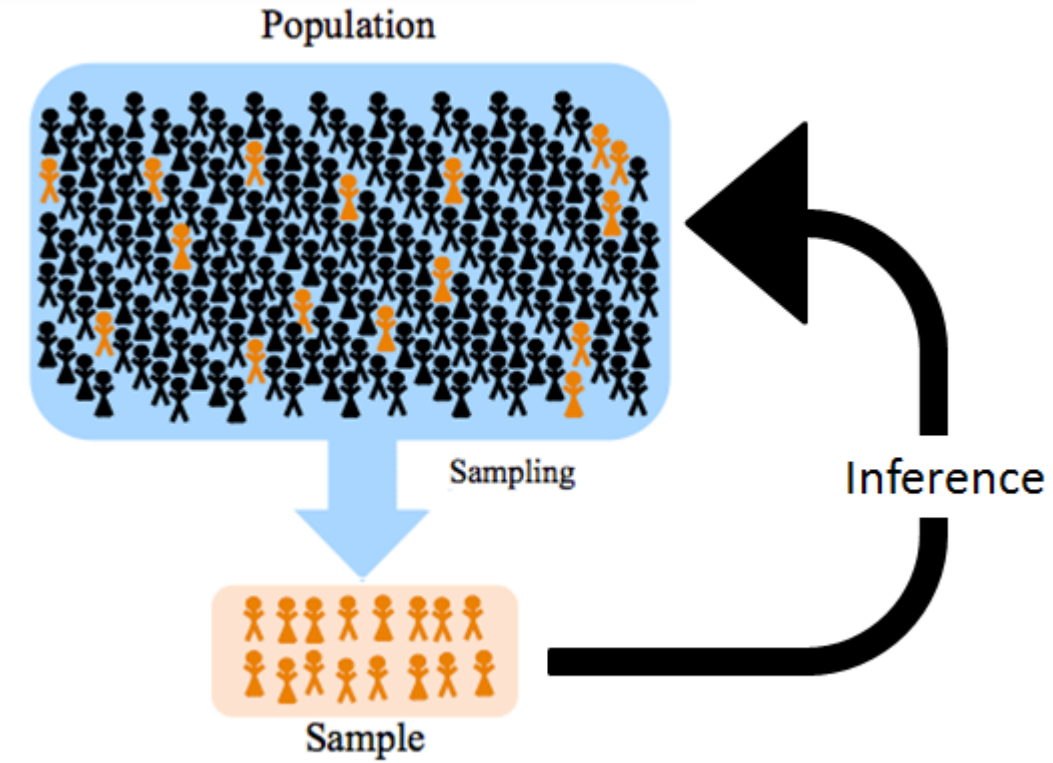
Sampling Distributions (example)

- For any estimate, the **sampling distribution** is the distribution of that estimate calculated for all possible samples of size *n* drawn from a population.



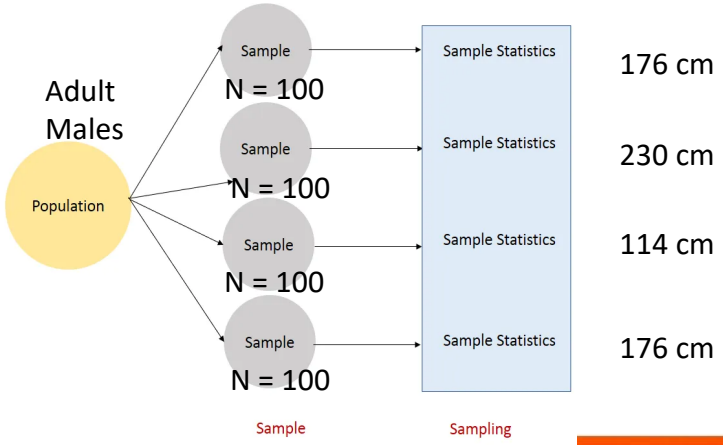
Sampling Distributions

- **Sampling distributions** can be used to make inferences about the true parameter of a population



Sampling Distributions

If you repeatedly take samples from a population, you will get a variety of different means, but some will be repeated.

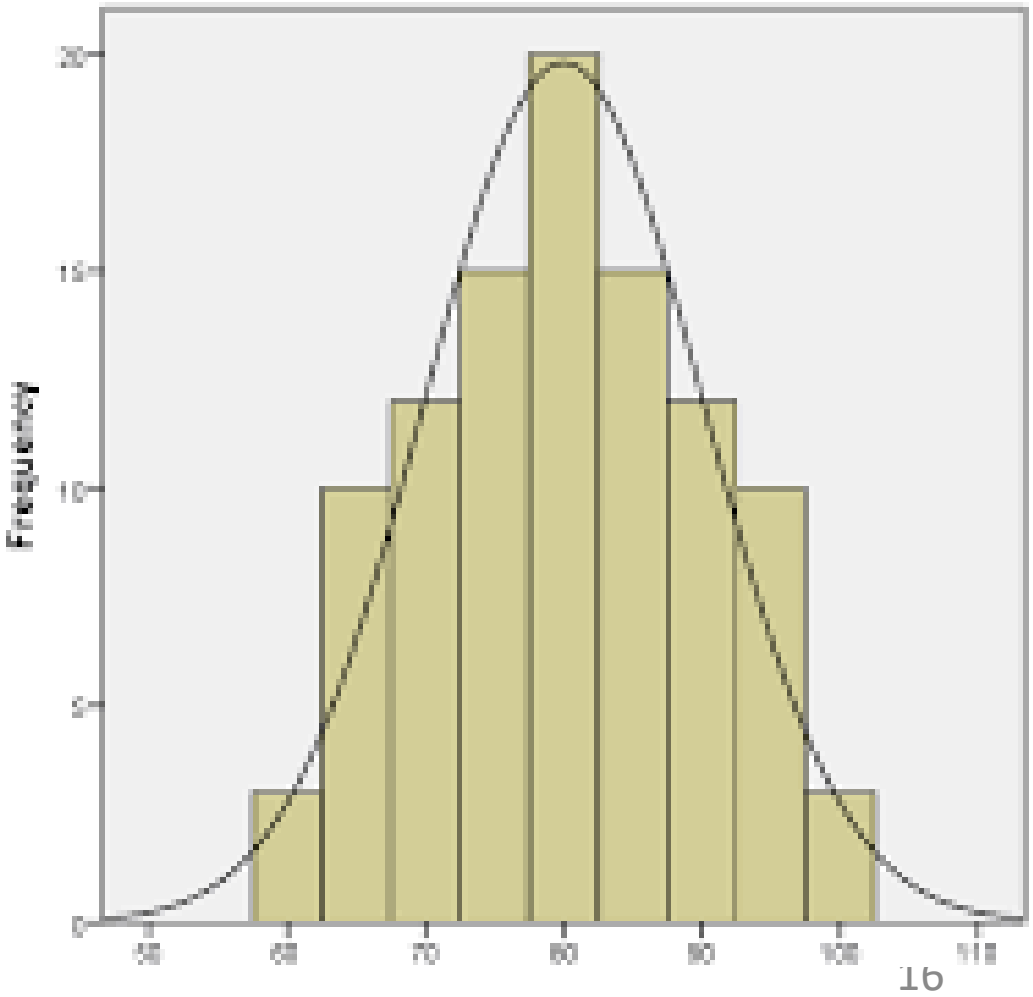


176 cm

230 cm

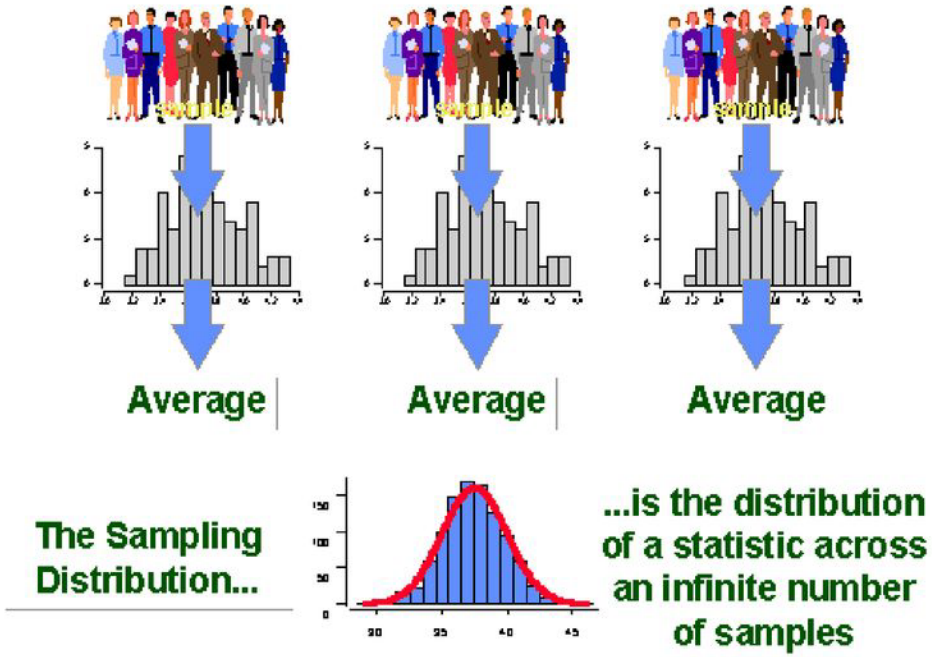
114 cm

176 cm



Sampling Distributions

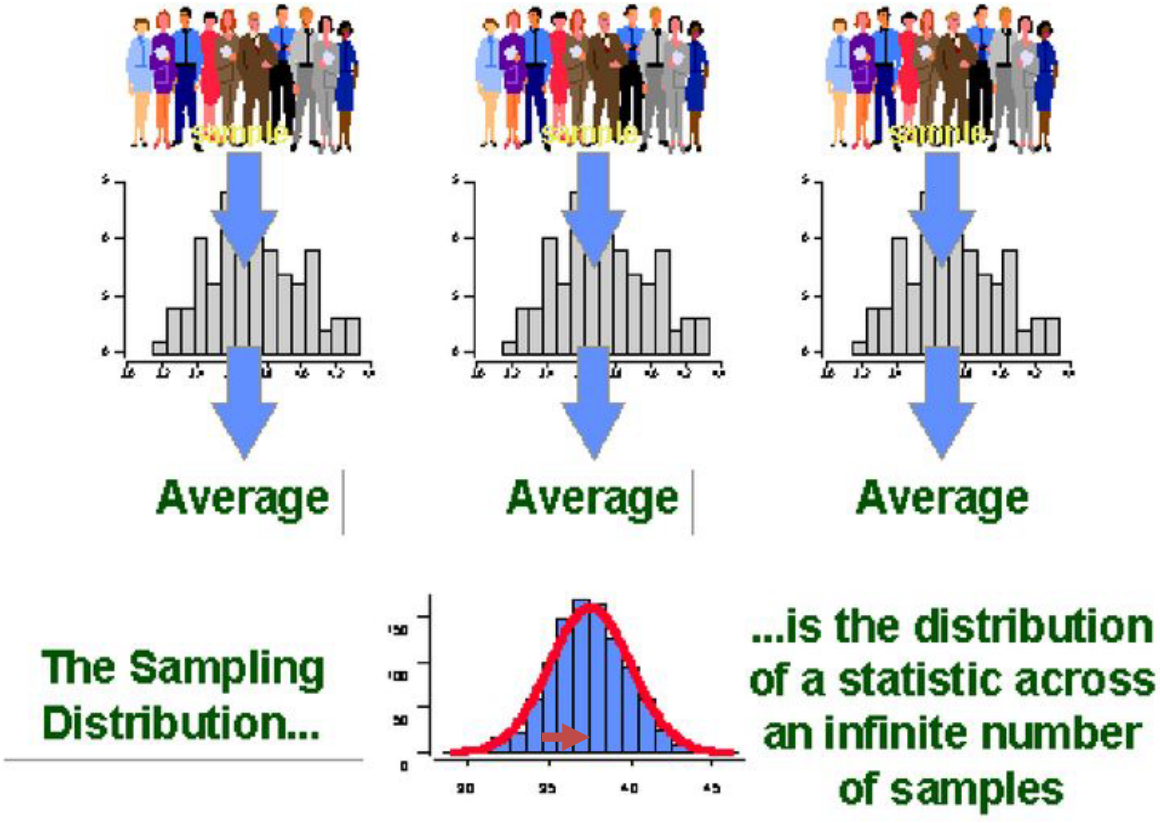
- If we repeatedly calculate the mean of randomly drawn samples from a population, and then collect them all together, we get a **sampling distribution**, which we can use to estimate the population mean.



http://www.socialresearchmethods.net/kb/sampstat.php

Sampling Distributions

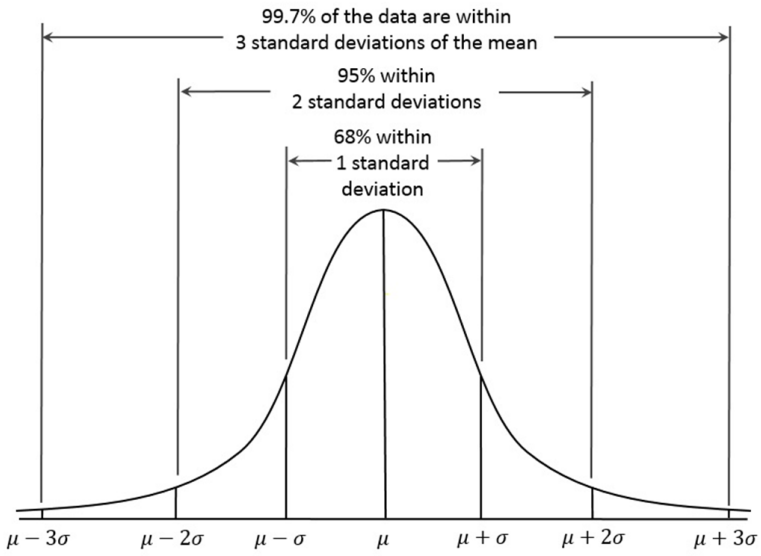
- You might be wondering, how will we know where the mean we collected falls within the sampling distribution?



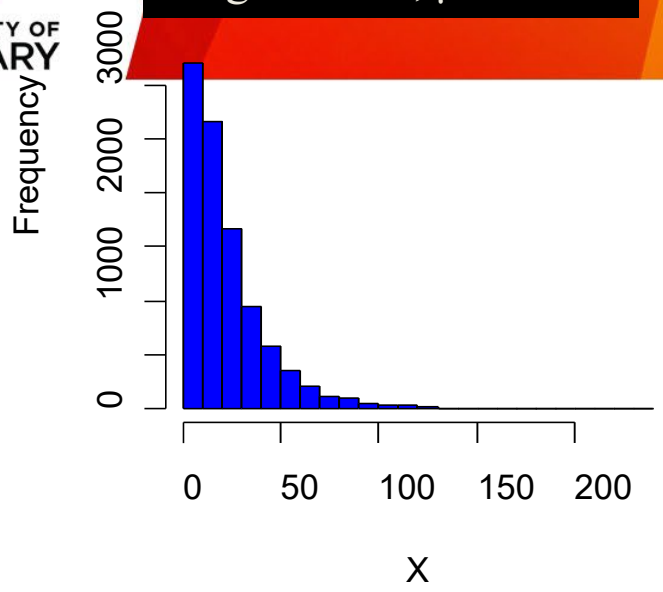
http://www.socialresearchmethods.net/kb/sampstat.php

The sampling distribution is ‘normal’

- The normal distribution
  - **Central limit theorem:** Means of random variables tend to be normally distributed.
  - As the number of observations in sample approaches infinity, sampling distribution of the mean will approach the normal distribution.
  - Great, but what does ‘normal’ mean?



Original data, μ=22.2



Sampling Distributions

Implication:

So we know that we can pretty accurately estimate a population parameter using a sample, as long as the sample is 'big enough'

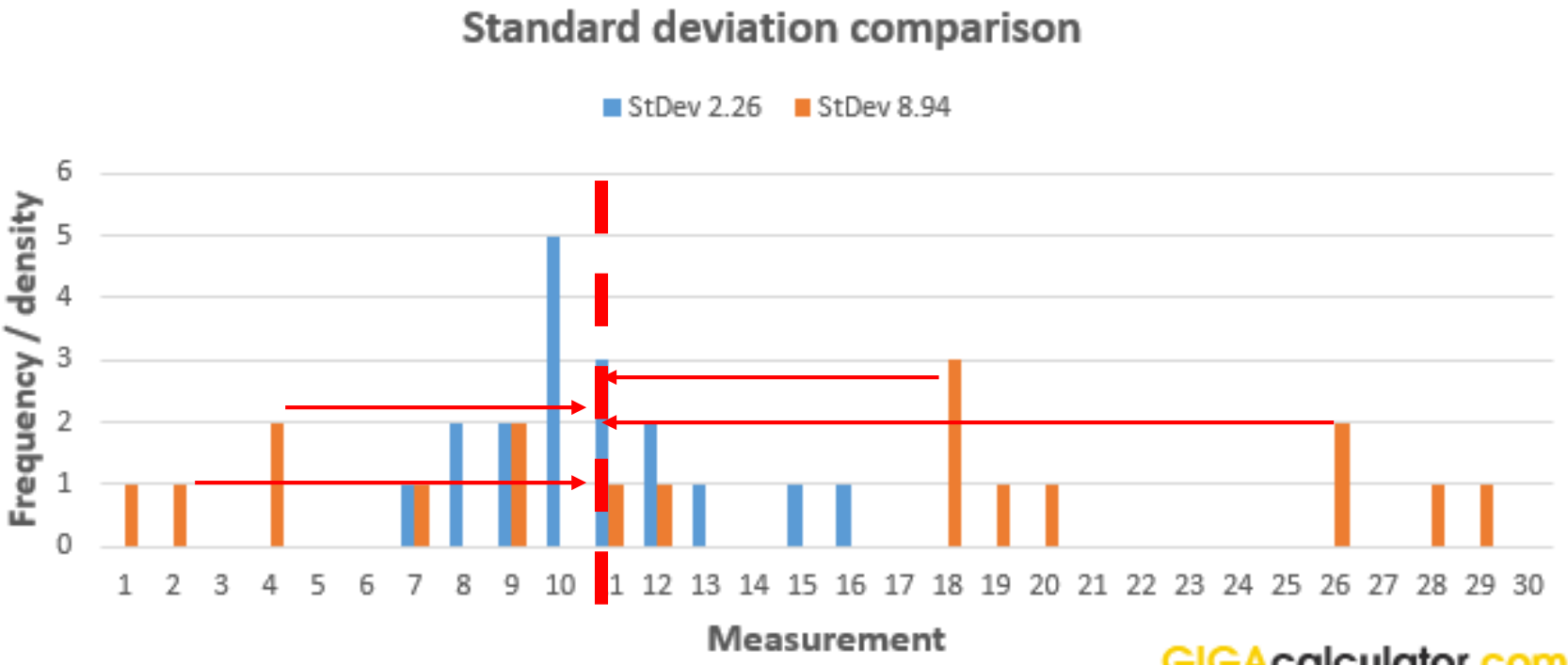
Practical Application:

Larger sample sizes help you accurately estimate what is going on in the population.

What is 'big enough'? In theory, about 30 should get you there, but in practice 100 is generally considered big enough.

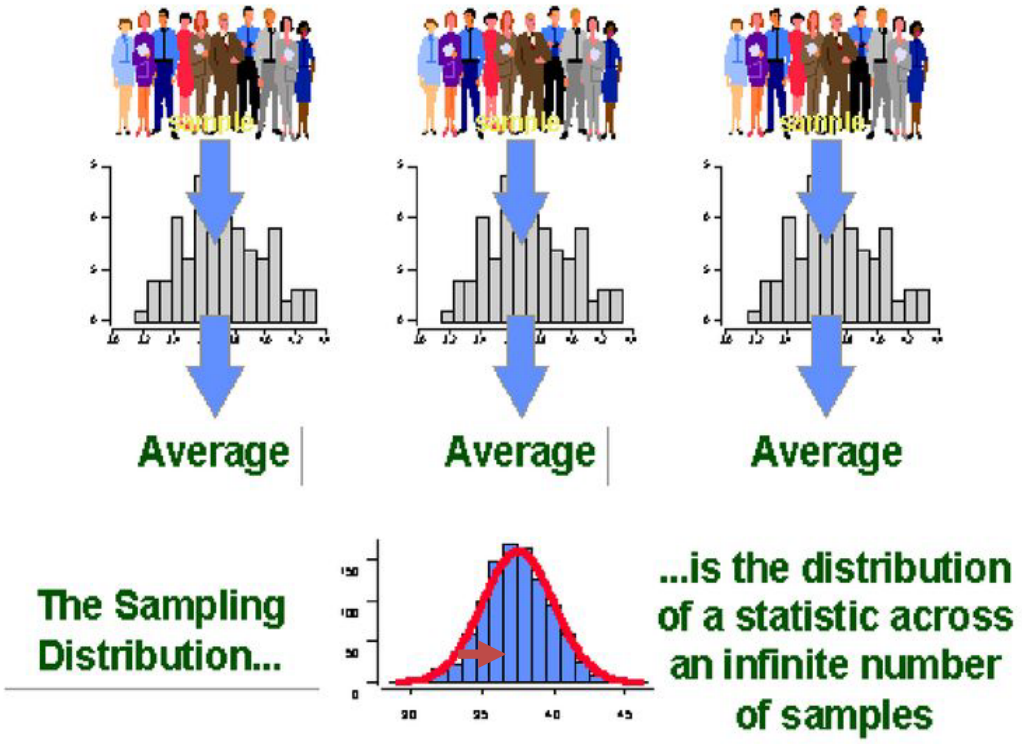
Estimating Error

- We can calculate the variability within our sample by adding up the differences between the individual values and the sample mean. This is called the standard deviation.



Estimating Error

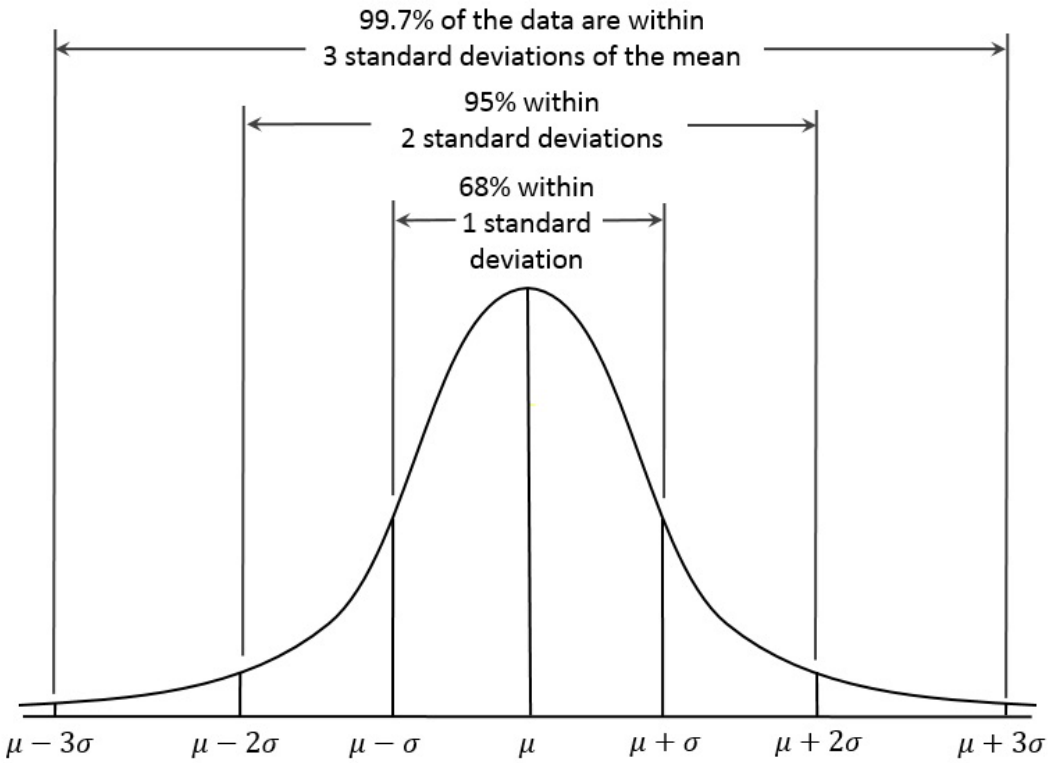
- We can also calculate the variability within our sampling distribution. That is, we calculate all the deviations of the sample means from the population mean. This is called the standard error.



http://www.socialresearchmethods.net/kb/sampstat.php

Standard Error

- The **standard error of the mean** (SE) is the standard deviation of **these** sample means
- **Sampling distributions** can be used to make inferences about the true mean of a population



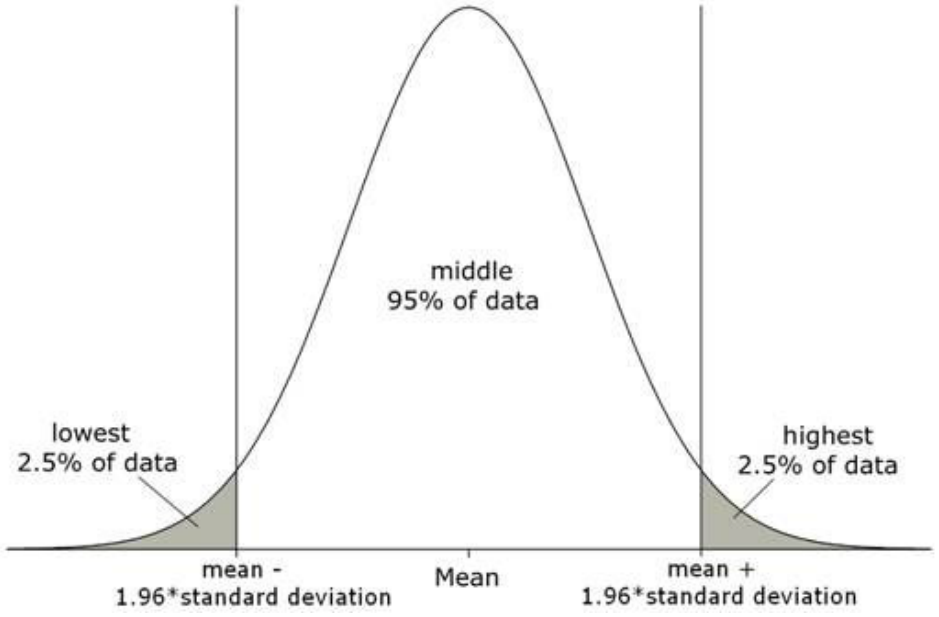
UNIVERSITY OF CALGARY

Inferences about the mean

- Usually standard error (of the population) is unknown
- Use SD of sample (S) as estimate of population SE
- With sample S, estimate the SE as SE = SD / $\sqrt{n}$
- As *n* increases SE decreases
- This means that if we use our sample mean as an estimate of the population mean, our estimate will be more accurate as our sample size increases

Confidence Intervals

- Given the normal sampling distribution of the means, 95% of sample means will fall within 1.96 standard errors of the sample mean.
- Because there is error in our estimate of the mean, we can use the normal distribution to create a **confidence interval**.



Confidence Intervals

- Example: A researcher reports a mean systolic blood pressure of 129 with a 95% confidence interval of [125 - 133]. What does this mean?

  - Because of the properties of the normal distribution, we are confident that the true mean systolic blood pressure is between 125 and 133.
  - In fact, there is only a 5% chance that our CI does not contain the true population mean.
  - Confidence intervals give a sense of the precision of the estimate.
    - Wide confidence intervals indicate a less precise estimate than narrow CIs.
  - How did we obtain the CI?
    - We used the SD of our sample to estimate the SE of the population (~2 in this case).
    - We multiplied the SE by 1.96 (95% interval)
    - We added this 'error' (+ or - 4 in this case) to the mean.

Confidence Intervals

- Example: A researcher reports a mean systolic blood pressure of 129 with a 95% confidence interval of [125 - 133]. What does this mean?

  - How did we obtain the CI?
    - We used the SD of our sample to estimate the SE of the population.
      - SE = 20/ $\sqrt{}$ 100 = 2
    - We multiplied the SE by 1.96 (95% interval)
      - 2*1.96 = ~4
    - We added this 'error' to the mean.
      - 129+4 = 133; 129-4=125

Material Covered in the Biostatistics Lectures

- Lecture 1B- Inference
  - The logic and use of statistics
  - Sampling Distributions
    - The normal distribution
    - Confidence intervals
  - Estimation and Hypothesis testing
    - Difference of Means
  - Power and Sample Size

UNIVERSITY OF CALGARY

Hypothesis Testing: Difference in means

- Interested in comparisons between groups or associations among variables.
  - Does the BP of 2 groups actually differ?
  - Is folate really associated with neural tube defects?
- Humans vary in their response to treatments, exposures etc.
- How do we know that observed effects can be *attributed* to the “treatment” or “grouping”?

Hypothesis Testing: Difference in means

- When comparing two measures of interest (e.g. means), we compare the observed difference in these to the distribution of differences we would expect to occur under the assumptions of **null hypothesis**.

- **H_0 :** We refer to the hypothesis of no difference or no effect as the **null hypothesis.**
- **H_A :** We specify an **alternative hypothesis.**

Example-Two independent groups

- Researchers want to compare the 24-hour total energy intake in two populations using two samples: lean and obese people
- Let μ_L = mean energy intake for the population of lean people
- Let μ_O= mean energy intake for the population of obese people
- H_0: $\mu_L = \mu_O$
- H_A: $\mu_L \neq \mu_O$

Hypothesis tests

- Most of the tests we will see in this course will have the form of:

$$\text{test statistic} = \frac{Observed - hypothesized}{\text{standard error of observed}}$$

- *"observed"* comes from sample, $\bar{x}$, (sample mean), or differences of means between groups.
- *"Standard error of the observed"* will take different formulations but will still refer to variability in the estimation of the parameter of interest (estimated from our sample).

Hypothesis tests

- Most of the tests we will see in this course will have the form of:

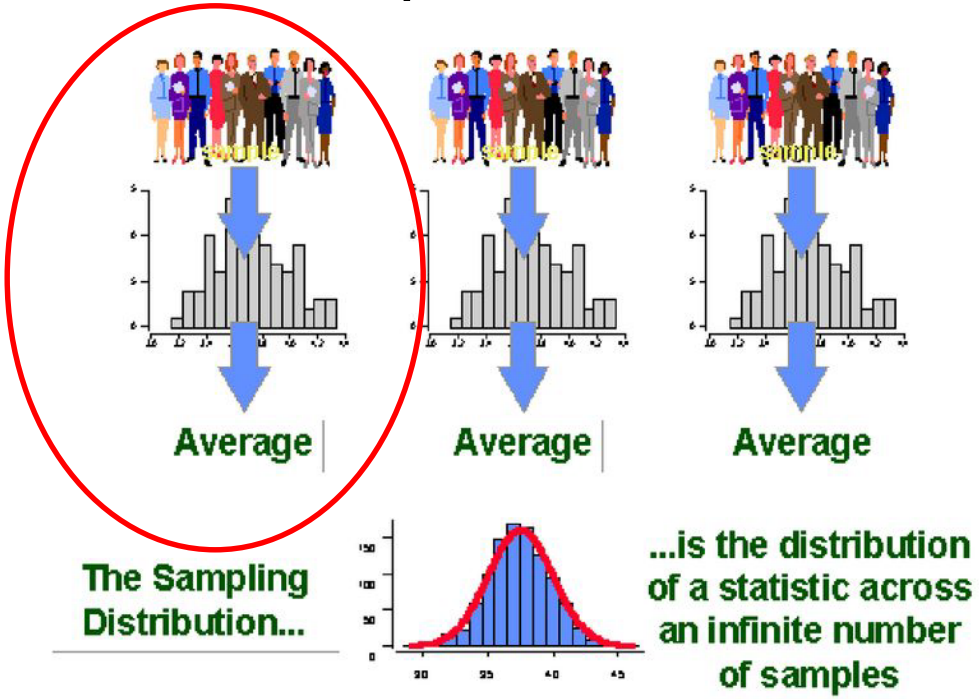
$$\text{test statistic} = \frac{Observed - hypothesized}{\text{standard error of observed}}$$

- *"hypothesized"* – here we are going to do a little trick to make the problem tractable.
- By making the null our hypothesized result, it is by definition zero, and we can drop this part of the equation.

$$\text{test statistic} = \frac{Observed}{\text{standard error of observed}}$$

t-tests cont…

- Because we don't know the population standard deviation (the standard error), we estimate the population standard deviation using data from the sample.



$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}$$

http://www.socialresearchmethods.net/kb/sampstat.php

Example: 24h energy expenditure

24 hour total energy expenditure (MJ/day) in groups of lean and obese people (Prentice et al, 1986)

	Lean (n=13)	Obese (n=9)
	6.13	8.79
	7.05	9.19
	7.48	9.21
	7.48	9.68
	7.53	9.69
	7.58	9.97
	7.90	11.51
	8.08	11.85
	8.09	12.79
	8.11	
	8.40	
	10.15	
	10.88	
Mean	**8.07**	**10.30**
SD	**1.24**	**1.40**

Data from Altman, 1995, Practical Statistics for Medical Research, Table 9.4, p. 193

Independent groups t-test

- H_0: True mean of Lean (μ_L) = True mean of obese (μ_O)
- H_A: True mean of Lean (μ_L) ≠ True mean of obese (μ_O)

$$t_{(n_O + n_L - 2)} = \frac{\text{Observed difference} - \text{Hypothesized difference}}{\text{Error}}$$

Observed difference: 10.3 - 8.07 = 2.23
Hypothesized difference: 0
Standard Error = 0.5656 (pooled from the 2 samples)

- t = 2.23 – 0/0.5656 = 3.94
- Compare *our obtained value of t* to the **critical value** of t (the value of t associated with our alpha level (two-tailed)).

t-tests cont…

- We can then use the t-distribution to make inferences about differences between our sample means.

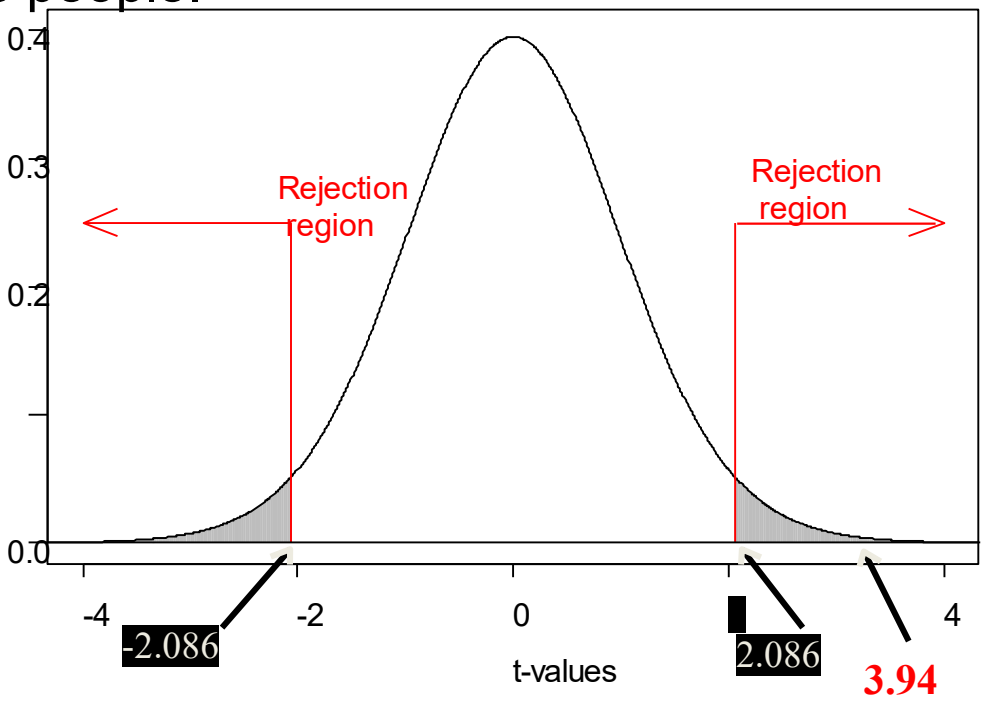
Critical values of *t* for two-tailed tests

Significance level (α)

Degrees of freedom (*df*)	0.2	0.15	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	4.165	6.314	12.706	25.452	63.657	127.321	636.619	1273.239
2	1.886	2.282	2.920	4.303	6.205	9.925	14.089	31.599	44.705
3	1.638	1.924	2.353	3.182	4.177	5.841	7.453	12.924	16.326
4	1.533	1.778	2.132	2.776	3.495	4.604	5.598	8.610	10.306
5	1.476	1.699	2.015	2.571	3.163	4.032	4.773	6.869	7.976
6	1.440	1.650	1.943	2.447	2.969	3.707	4.317	5.959	6.788
7	1.415	1.617	1.895	2.365	2.841	3.499	4.029	5.408	6.082
8	1.397	1.592	1.860	2.306	2.752	3.355	3.833	5.041	5.617
9	1.383	1.574	1.833	2.262	2.685	3.250	3.690	4.781	5.291
10	1.372	1.559	1.812	2.228	2.634	3.169	3.581	4.587	5.049
11	1.363	1.548	1.796	2.201	2.593	3.106	3.497	4.437	4.863
12	1.356	1.538	1.782	2.179	2.560	3.055	3.428	4.318	4.716
13	1.350	1.530	1.771	2.160	2.533	3.012	3.372	4.221	4.597
14	1.345	1.523	1.761	2.145	2.510	2.977	3.326	4.140	4.499
15	1.341	1.517	1.753	2.131	2.490	2.947	3.286	4.073	4.417
16	1.337	1.512	1.746	2.120	2.473	2.921	3.252	4.015	4.346
17	1.333	1.508	1.740	2.110	2.458	2.898	3.222	3.965	4.286
18	1.330	1.504	1.734	2.101	2.445	2.878	3.197	3.922	4.233
19	1.328	1.500	1.729	2.093	2.433	2.861	3.174	3.883	4.187
20	1.325	1.497	1.725	2.086	2.423	2.845	3.153	3.850	4.146

Independent groups t-test (cont)

- Values above 2.086 are unlikely in the t distribution when there are 20 df.
- Conclusion: Reject the null hypothesis of no difference in energy expenditure. Instead, we believe there is evidence to suggest a difference in the 24-hour total energy expenditure between lean and obese people.
- There is less than 5% chance that we would have obtained this difference in energy expenditure if there were no differences between lean and obese people.



Confidence Limits

- Point estimates are great (simple) but you should always report the range of estimates that are plausible, given your data.
- This requires creating a confidence interval (CI).

- 95% CI = sample $mean_O$ – sample $mean_L$ ± ($t_{(critical)}$*$SE_{difference}$)
  - For a 95% CI the critical value is 2.086
  - The mean difference between groups is 2.23
  - The SE of the difference is 0.5656

- 2.23 ± 2.086*0.5656
- 95% CI is 1.05 to 3.41

Confidence Limits

- Interpretation: The likely range of the true difference in population can be as small as 1.05 or as big as 3.41.
- That is, that the mean energy expenditure for obese is at least 1.05 units larger and up to 3.41 units larger
- Relationship between statistical test at alpha level of 5% and 95% confidence interval
  - A CI that includes 0 has a p-value > 0.05

UNIVERSITY OF CALGARY

Advice

- You aren't expected to know a lot of detail about statistical tests.
- Familiarize yourself with the basics of hypothesis testing…this will be very helpful for your understanding of literature and a bit of effort will serve you well in the future.
- If you have a significant p value – you reject the null, which means that **there is reason to doubt that the true effect is zero (there is a 'real' difference).**
- If you have a non-significant p value – you fail to reject the null, which means that you **cannot distinguish the effect from zero.**
- **Report confidence intervals!**

Material Covered in the Biostatistics Lectures

- Lecture 1B- Inference
  - The logic and use of statistics
  - Sampling Distributions
    - The normal distribution
  - Estimation and Hypothesis testing
    - Confidence intervals
    - Difference of Means
    - Types of error
  - Power and Sample Size

UNIVERSITY OF CALGARY

Statistical Power

- **Statistical power (1-beta)** is the probability of rejecting the null hypothesis when it is false in the population
  - (i.e. it is the probability of correctly concluding there is evidence of a difference).
- Studies with small sample sizes may not have adequate power to detect clinically meaningful effects.
  - These studies are sometimes referred to as being **underpowered**.
  - What is the difference between clinical and statistical significance?

Sample Size/Power Calculations:

- Background: What are they and why do them?
- Typically we are interested in calculating a sample size for a study that is ‘adequately powered’.

- Adequate is usually defined as a power of at least 80%.

Sample Size Estimates: Inference

- Inference
  - In research, can't measure everyone.
  - Make inferences on "true" or underlying characteristics of a population on the basis of data collected from a sample.
  - The more subjects measured, the more accurate our estimates will be. (if we sampled appropriately)
  - Measure too many, waste resources.
  - Measure too few, won't be able to detect effects of interest.
  - But how many patients?

UNIVERSITY OF CALGARY

How Many Subjects?

- This depends on several things
  - The kind of statistical tests to be used.
  - Choice of study design.
  - The variability of the values in the population (both true differences and measurement error).
  - The alpha level (probability of concluding that there is an effect or difference, when in truth, there is no effect).
  - The statistical power (the probability that we will conclude there is an effect, given that the true effect is of a given size).
  - The magnitude of the hypothesized effect. (e.g. What is a clinically relevant difference in means?)
  - The number of drop-outs or non-responders.

Hypothesis based estimates cont…

- Power can only be specified for explicit alternative hypotheses

  e.g. H_A: $\mu_1 = 90$; $\mu_2 = 110$, or alternatively

  H_A: $\mu_2 - \mu_1 = 20$

- For hypothesis-based sample size calculations, have to state the size of the treatment effect that we wish to be able to detect.
- This difference is usually taken to be one that has clinical relevance.
  - (i.e. what is the smallest difference that would be considered clinically meaningful?).

Example: Comparing Means

- Researchers would like to do a clinical trial comparing methods of treating dorsal wrist pain. Treatments will be typical physiotherapy (control group) vs physiotherapy plus shock wave therapy (treatment group).
- They will use the PRWE, a validated measure with a maximum score of 100, to evaluate pain and function.
- Their outcome will be the change in the PRWE from the pre-treatment baseline score to the 6-month follow-up score.

What do we need for calculations?

- **Variability of the outcome**: Because they are studying change in PRWE, they need to know the variability of 6-month changes in the PRWE.
- From the medical literature and correspondence with the developer of the PRWE, they established that the range of the SD for change over time was between 17 and 22.

- **Clinically relevant improvement:** Based on other studies, the researchers propose that a clinically relevant difference in change would be 20 points.

- **Alpha Level:** 0.05
- **Power:** 80%
- **Dropout rates:** They anticipate a 10% dropout rate in each group.
- **Sample size calculators:**
  - http://www.cs.uiowa.edu/~rlenth/Power/

Note: This software was recently revised, and apparently some users are having trouble with it running properly. If that happens to you, here is a link to the old version If you do have problems, I'd appreciate your reporting to me (russell-lenth@uiowa edu) the following information: (1) which web browser you use and its version number; (2) what version ofJ ava you have installed.. One way to get the latter is to get to a terminal window (e.g. in Windows, [Start] / [Run...] "cmd"), and type the command `java -version` . Thank you

Java applets for power and sample size

Sele ct the analysis to he use din your study:

- CI for one proportion
- Test of one proportion
- Test comparing two proportions
- CI for one mean
- One-sample t test (or paired t)
- Two-sample t test (pooled or Satterthwaite)
- Linear regression
- Balanced ANOVA (any model)
- R-square (multiple correlation)
- Generic chi-square test
- Generic Poisson test
- Pilot study

Run selection

This software is intended to be useful in planning statistical studies. It is not intended to be used for analysis of data that have already been collected.

Each selection provides a graphical interface for studying the power of one or more tests. They include sliders (convertible to number-entry fields) for varying parameters, and a simple provision for graphing one variable against another.

Each dialog window also offers a Help menu Please read the Help menus before contacting me with questions.

The "Balanced ANOVA" selection provides another dialog with a list of several popular experimental designs, plus a provision for specifying your own model.

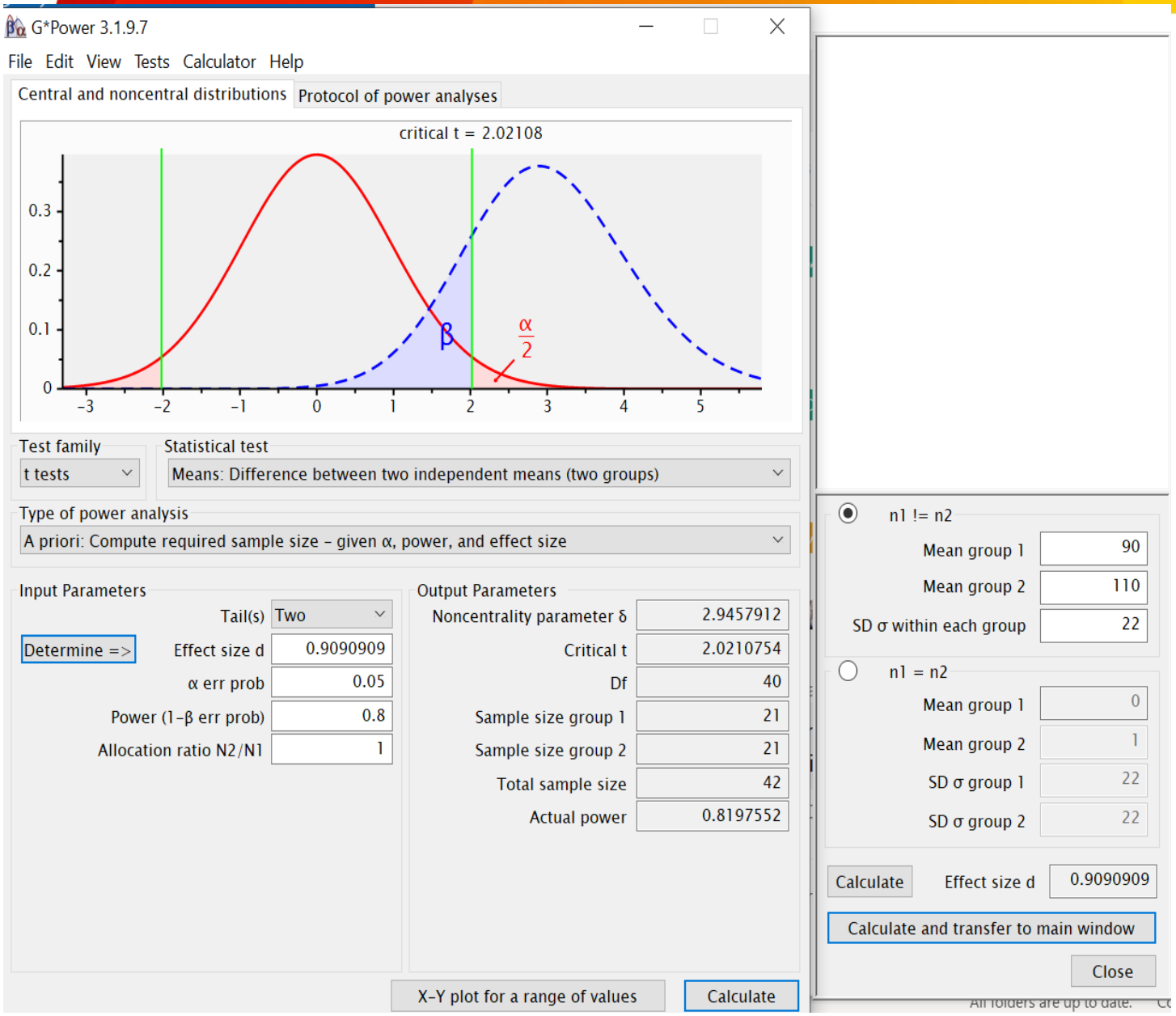
Note: The dialogs open in separate windows. If you're running this on an Apple Macintosh, the applets' menus are added to the *screen* menubar -- so, for example, you'll have two "Help" menus there!

You may also download this software to run it on your own PC.

Note: These require a web browser capable of running Java applets (version 1.1 or higher). If you do not see a selection list above, chances are that you either have disabled Java, your

Applet rvl.piface,apps,PiPicker started

G*Power sample size calculation

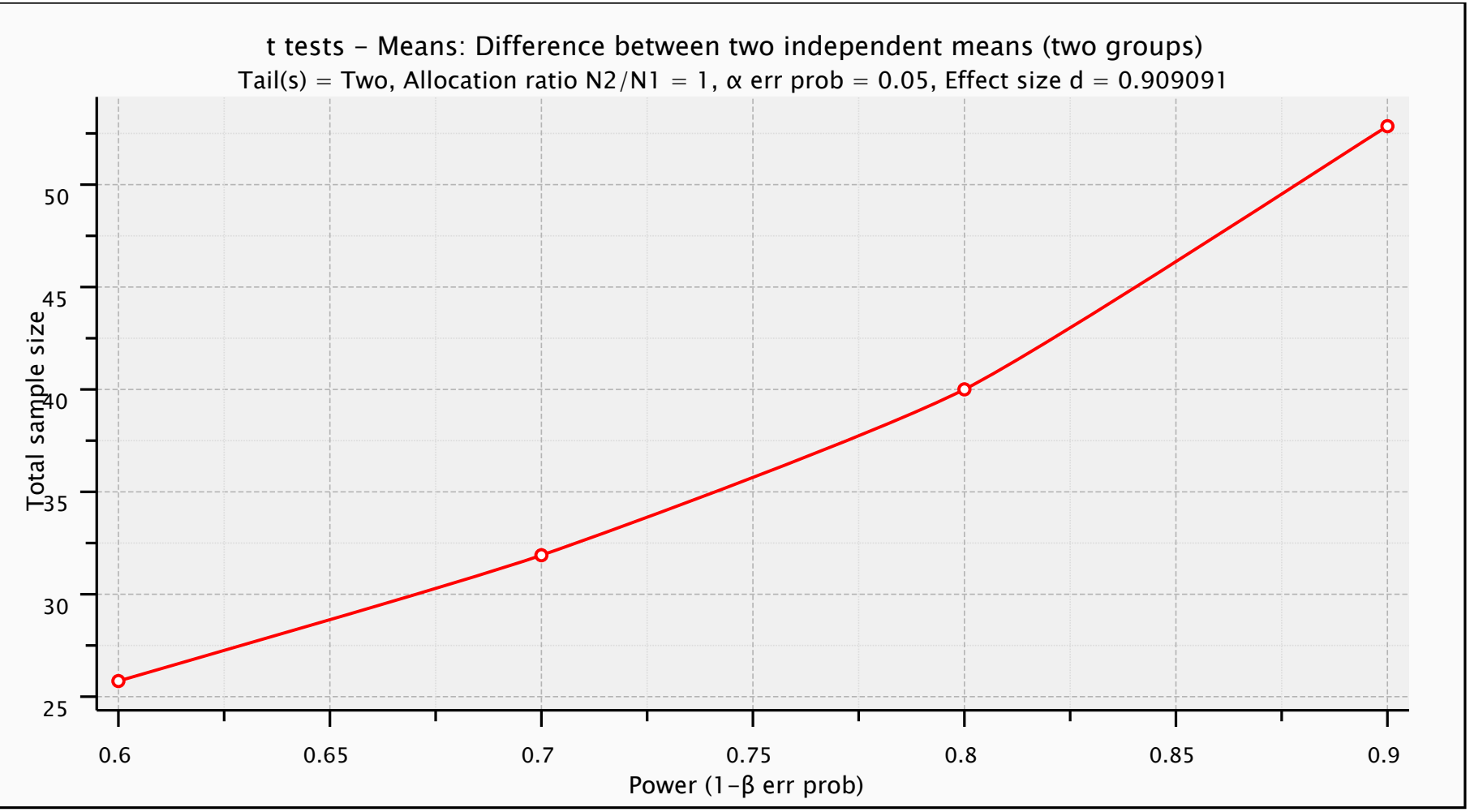


UNIVERSITY OF CALGARY

Example cont…

- From this example, will need 21 per group before dropouts.
- Expect 90% of patients to complete the trial, will need 21/.9 = 23 per group.

Effects of sample size on power



Some sample size resources

- G*Power (http://www.gpower.hhu.de/en.html) (free)
- PS Power (http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize) also free
- Russ Lenth's page: http://www.cs.uiowa.edu/~rlenth/Power/
- Within R-project for statistical computing, packages: power, pwr, Hmisc, EpiR, etc. (Free)
- Within Stata: power (and endless options)