

# Biostatistics: ~~A formal introduction~~ A love story

Raymond R. Balise, Ph.D.

Department of Health Research and Policy



Stanford / Packard Center for  
Translational Research in Medicine

# A Love™ Story...

- A biochemist, in a freak accident, has discovered a chemical that caused feelings of euphoria in her lab assistant. As a bench scientist, she knows exactly how to make the drug (which she decided to call Love™) but she has no idea how to investigate the efficacy of the drug.
- This is her story....

# Literature Reviews

- After doing extensive animal testing and determining tolerable doses in humans. She decides that depressed, sick people need Love™ but she does not know how to measure depression.
- 70,000 different questionnaires have been used to assess depression.
- She doesn't know which to choose!

# How to Measure Depression

- She calls a biostatistician who we will call i. I sends her to find a published test of depression and suggests she check the *Mental Measurements Yearbooks and Tests in Print*
- <http://www.unl.edu/buros/>
- In the Lane Medical Library reference room (Z5814.P8 M4) she finds the latest *Mental Measurements Yearbooks*.

# Mental Measurements Yearbook

- Purpose
- Population – ages it is intended for
- Publication dates
- Administration – group or individual
- Price
- Time to administer
- Author/Publisher
- Reviews

# Looking at Test Scores

- You want to get a pool of people to assess and see if there is any variability.
  - Sample from a population
  - Take a look at the observed scores
  - See what the most common values are
  - See what the extremes are
  - See if there is a pattern in scores
    - Do most people get a particular number or a number plus or minus a few points?

# Populations and Samples

- Define the population to whom you want to generalize.
- Sample people independently (or representatively) from that population.
  - Independent means that the chance of selecting one person does not impact the chance of getting another
  - If you sample people that are related you will need to attend to this important detail when analyzing your data
- Sampling people that are matched in some respects is not a bad thing but you need to consider it.
  - Get professional help....
- If you measure the same person repeatedly, it is not a bad thing.
  - Get professional help!

# Independence... why care? (1)

- There are at least two general types of statistics, parametric and non-parametric.
  - Parametric statistics say that data can be summarized using a couple of “parameters.”
    - You use parametric statistics to compare (among other things) mean values of some measure between two groups.
  - Non parametric statistics use ranks
    - You use non-parametric statistics to compare which group has higher overall scores for some measure comparing two groups.

# Independence... why care? <sub>(2)</sub>

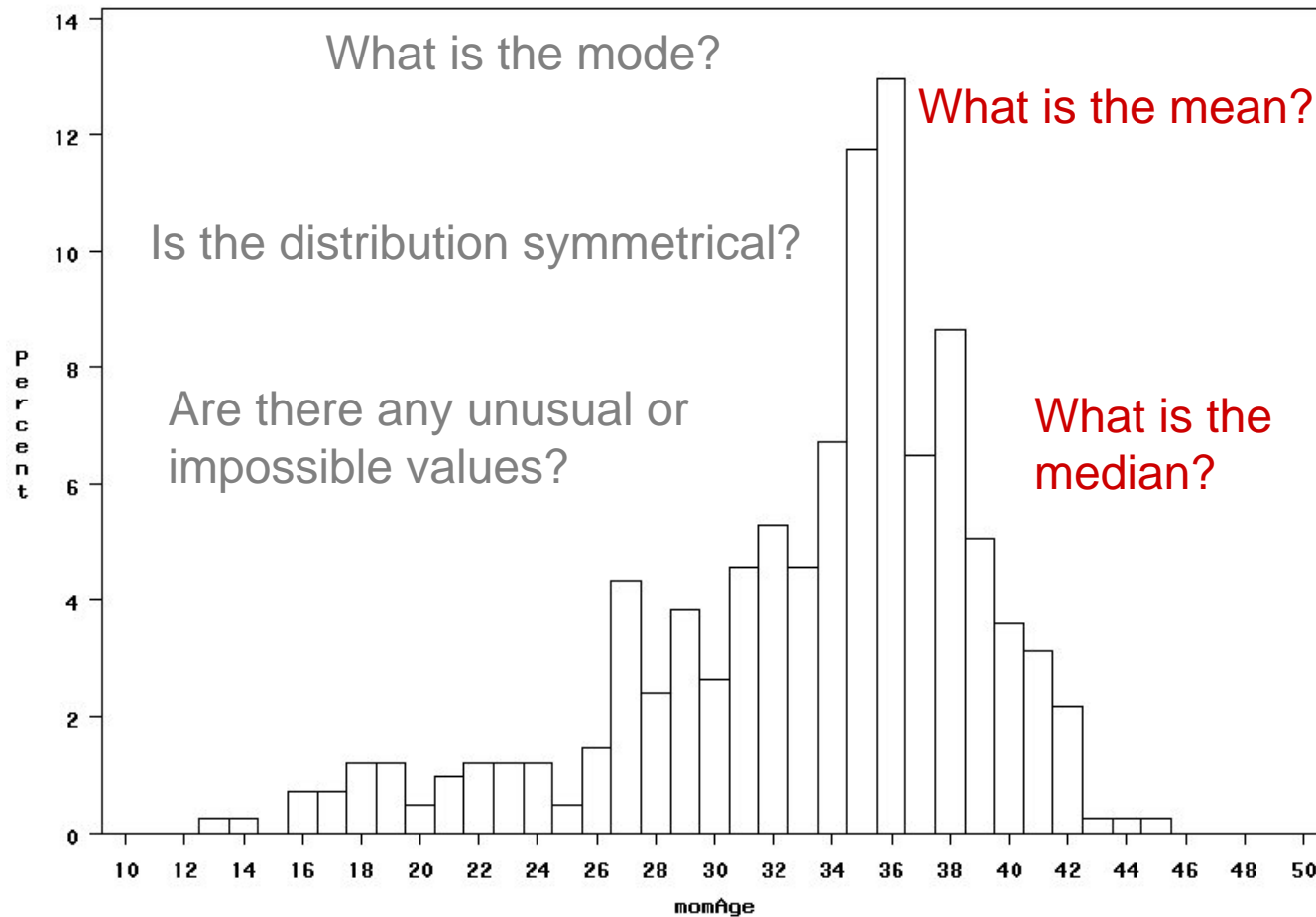
- Independence matters **a lot**
- All the common parametric statistics are done using two things: a measure of central tendency (mean) and a measure of variability (variance or standard deviation).
  - Your measure of variability will not match the population value if you have non-independent samples
    - Think about blood pressure measured on the same person 100 times vs 100 different people.
- Non-parametric statistics rank people from high to low. Are two people very low because they are siblings?

# Describing Data

- If you take a formal biostatics course, they will undoubtedly tell you how to hand calculate measures of central tendency and dispersion.
- If you ask me, I say just plot your data.
  - Histograms
  - Quantile plots
  - Boxplots

# Histogram

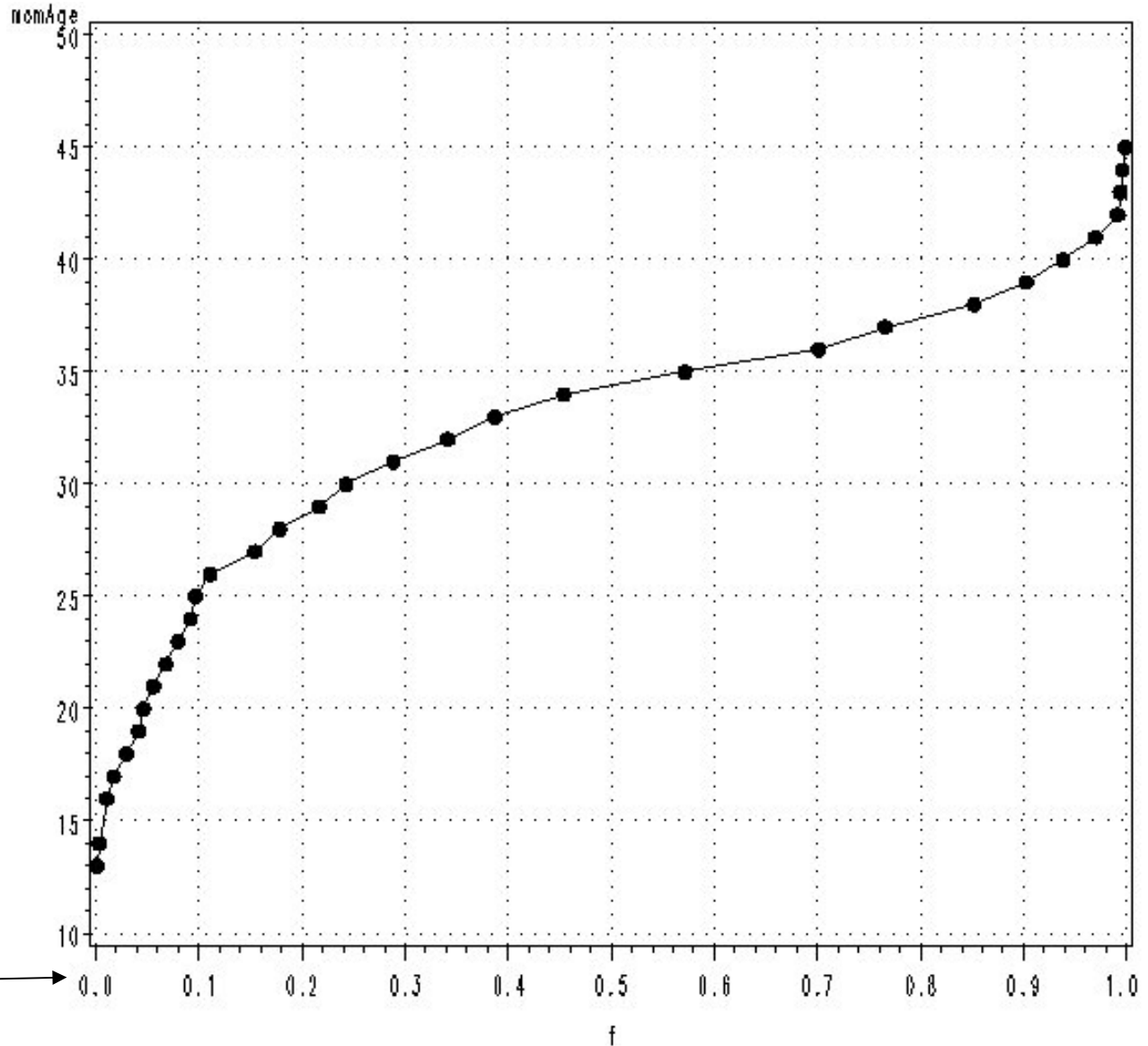
How many peaks?



Age at first birth in a sample of 560 mothers at a hypothetical hospital.

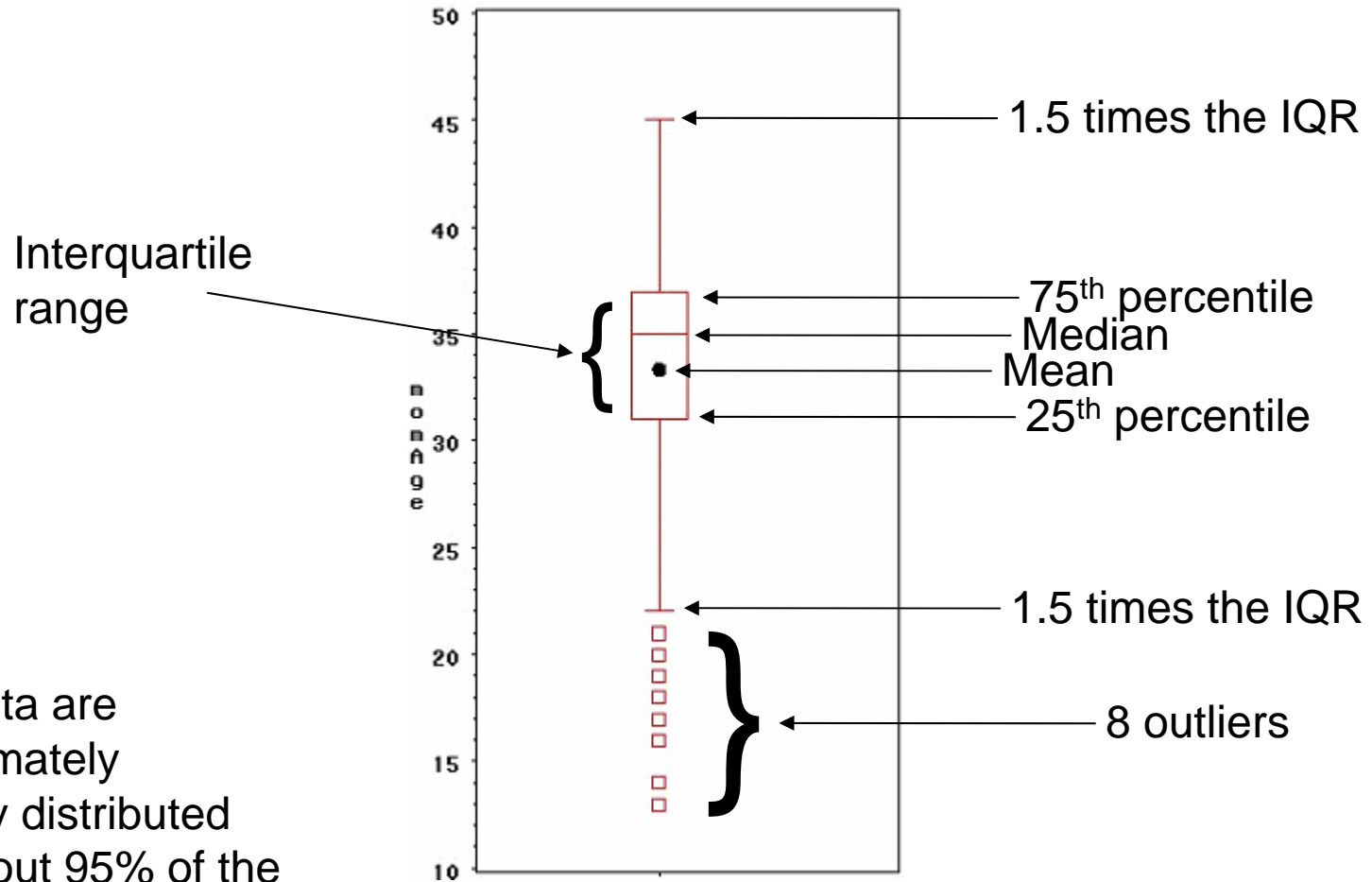
# Quantile Plot

Age at first birth in a sample of 560 mothers at a hypothetical hospital as quantiles.



Percentiles or multiply  
by 100 for %

# Boxplot



If the data are approximately normally distributed then about 95% of the data should fall within the 1.5 IQR range.

# Shape of a Distribution

- Does the histogram have a normal (aka Gaussian) distribution (bell shaped curve). If you plot a histogram of a biological outcome that is driven by many causal factors, it will typically look sort of like a bell shaped curve.
  - Age at childbirth
- Not everything is well described by a Gaussian distribution.
  - Age at leukemia: the histogram has two bumps (bimodal), one for childhood leukemia the other for later onset.
  - Income has a lower bound but the upper tail can go very high (positive skew).

# Is it really normally distributed?

- There are formal tests of normality.
  - Shapiro-Wilk test
- There are graphical ways to look to see if data is normally distributed.
  - Quantile Quantile plot (QQ plot)

# Why bother about the distribution?

- Statisticians are lazy... if you know your data is close to normally distributed, you can easily figure out how unusual a value is.
- You need to know two numbers (two parameters) to describe a normal distribution: the arithmetic mean and standard deviation.
  - The mean is just the sum of the values divided by the number of observations.
  - The standard deviation is calculated by taking the difference between each person's score and the mean, squaring (number times itself) each of the differences, adding up the squared differences and then dividing the sum by the number of observations (minus 1).

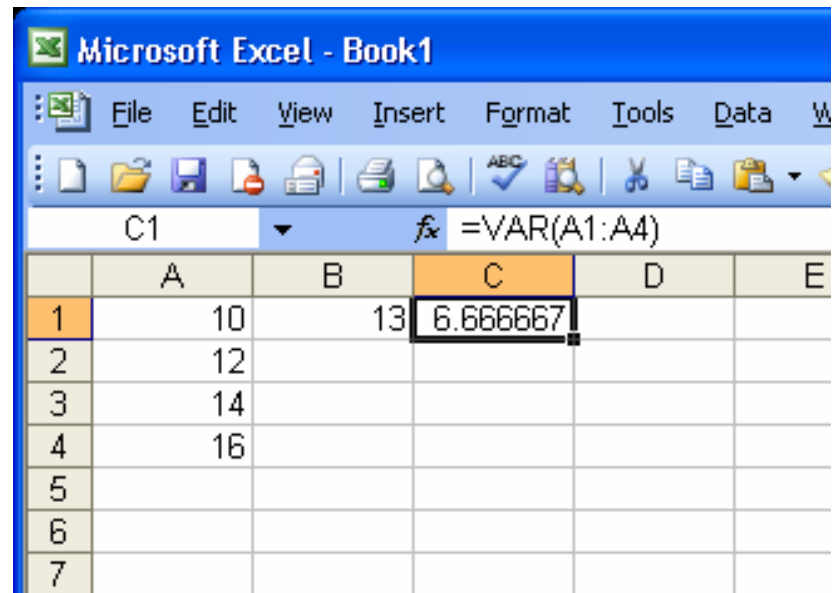
# Simple Descriptive Stats

- Nobody in their right mind calculates these things by hand. If you use software make sure it has the N-1 in the denominator. In Excel you can get a mean

=average()

or variance

=var()



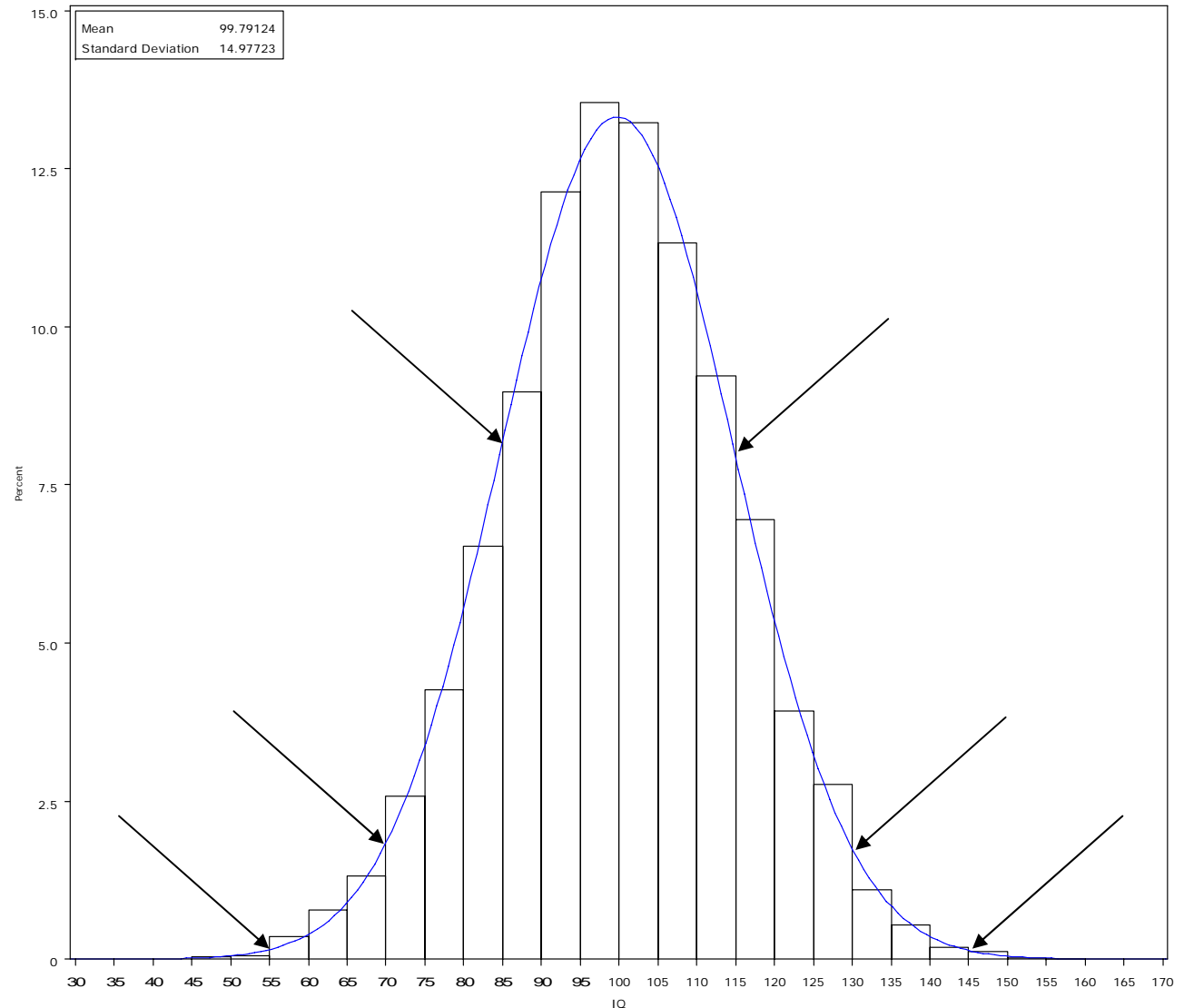
The screenshot shows a Microsoft Excel window titled "Microsoft Excel - Book1". The menu bar includes File, Edit, View, Insert, Format, Tools, Data, and Windows. The toolbar contains various icons for file operations and editing. The active cell is C1, and the formula bar shows the formula `=VAR(A1:A4)`. The spreadsheet data is as follows:

	A	B	C	D	E
1	10	13	6.666667		
2	12				
3	14				
4	16				
5					
6					
7					

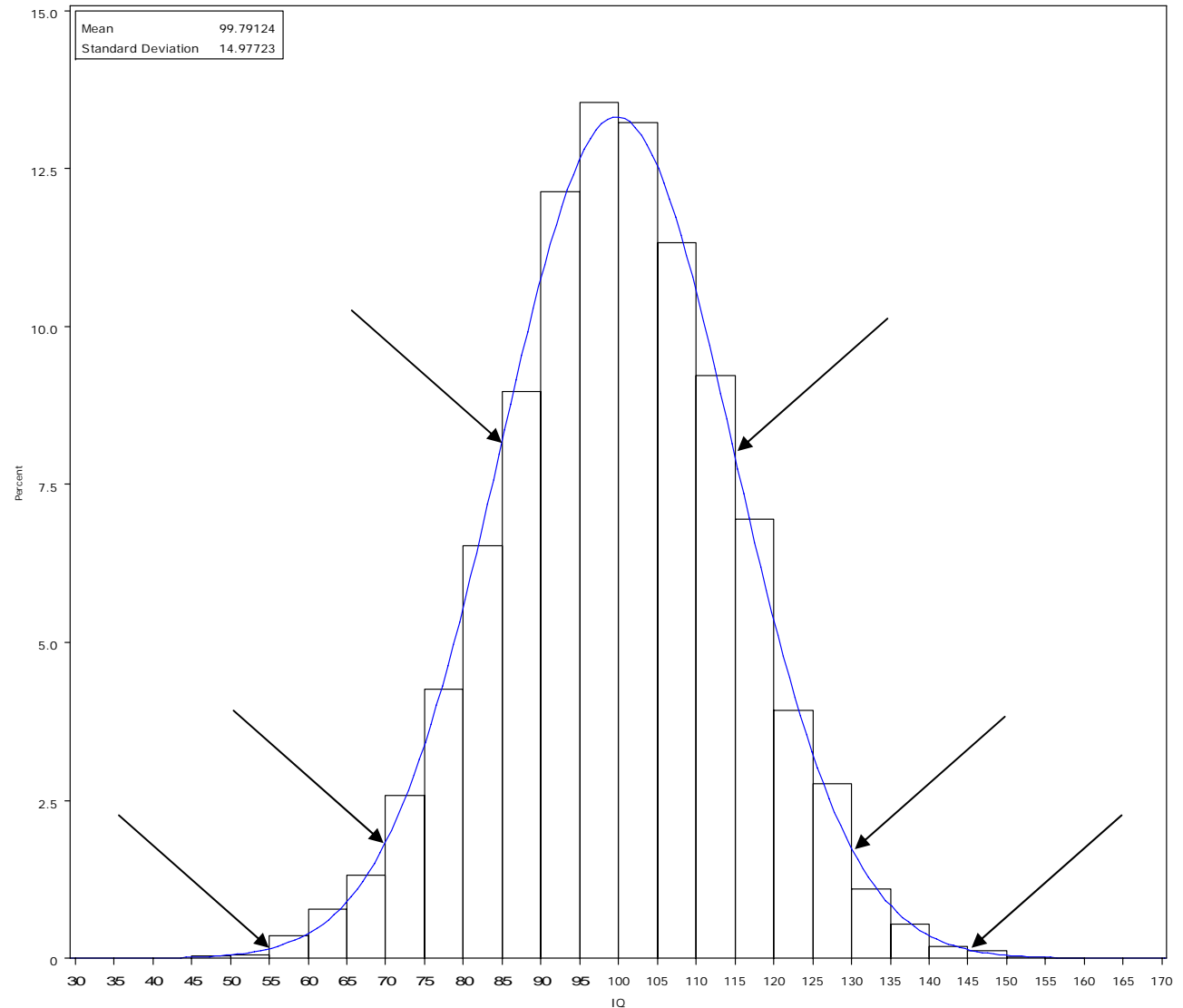
# Things are not normal

- When are means and standard deviations a poor choice to describe data?
  - If you have outliers
  - If you have highly skewed data
  - If you have more than one mode
- Other things (distributions) should not be shaped like a bell curve.
  - Risk for death
    - Low walking into the surgery, very high on the operating table but it drops precipitously after surgery and remains low
    - Hazard of death once infected with HIV

- Values are close to the mean and things far away are weird.
- If you draw a normal distribution, the places where the curve goes from convex to concave (and back to convex again are the standard deviations)



- If you add and subtract:
  - 1 SD from the mean, you get about 64% of the values
  - 2 SD (really 1.96 SDs) from the mean, you get about 95% of the values
  - 3 SD from the mean, you get about 99% of the values



# When data is not normal

- When your data is not well described by a bell shaped curve, you can “transform it” or use different shapes to describe and analyze the data.
  - These are **parametric** approaches to describe your data
- If you can't describe the data with a few numbers, you can usually look at the percentiles of the data and use rank orders to describe and analyze what is unusual.
  - These are **nonparametric** approaches to describe your data

# Components of a Standardized Test

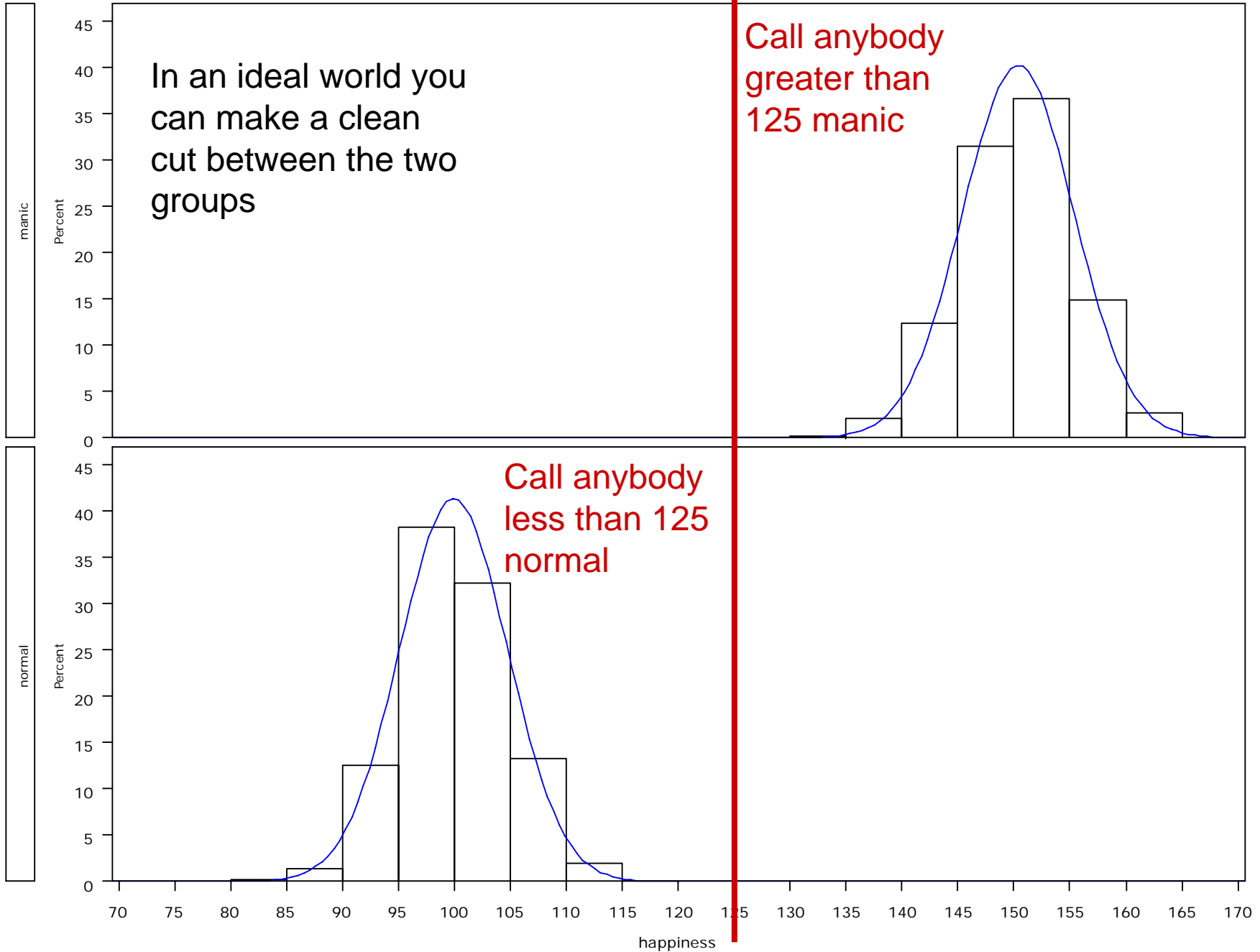
- Meanwhile back at the story at hand...
- In the index she finds about a dozen published tests that assess depression and decides to look at the Beck FastScreen for Medical Patients. Two reviewers have written articles which describe the instrument and how to score it. While well written, the reviews mention a couple of new terms.
  - Sensitivity
  - Specificity
  - Correlation

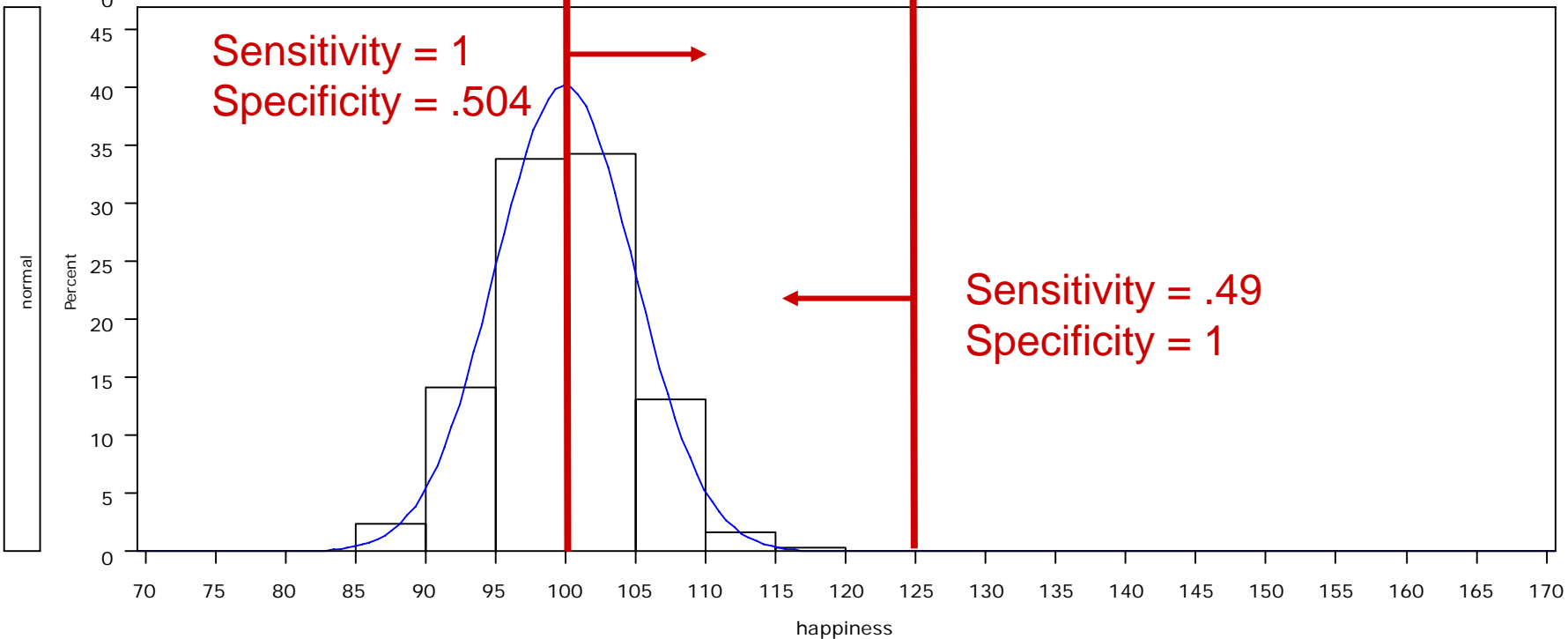
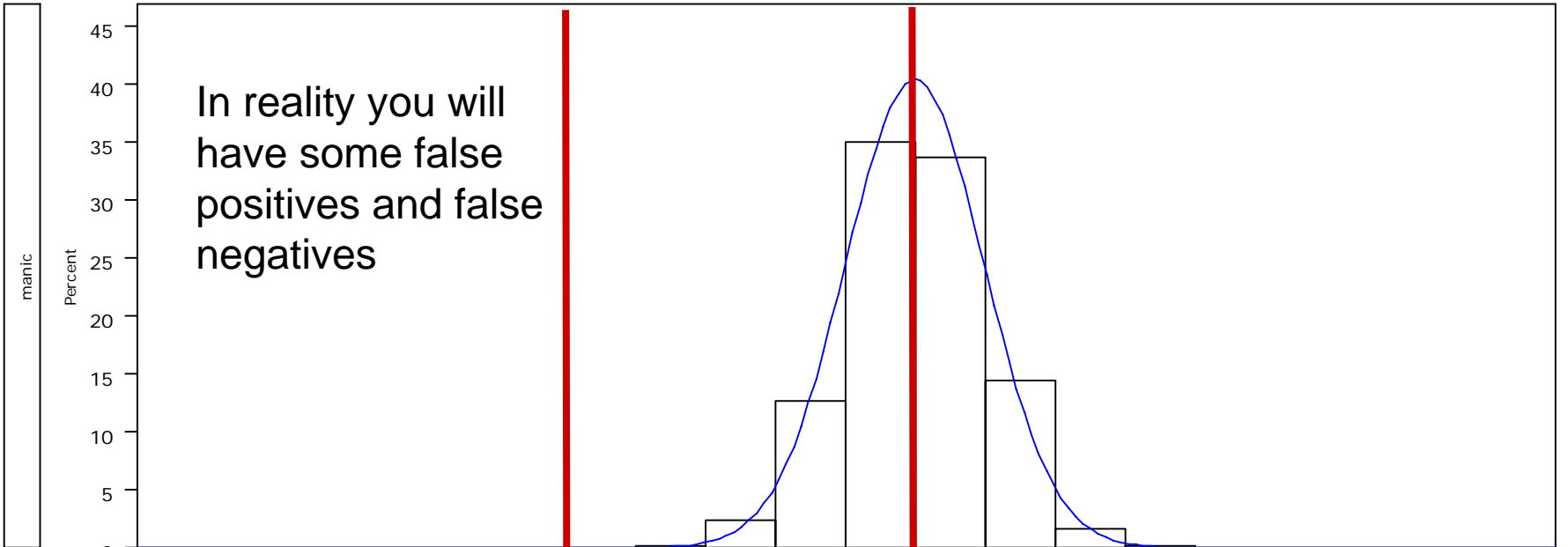
# Sensitivity and Specificity

- The sensitivity of a test is the ability of the test to correctly identify a subject who has a condition.
  - True Positive / (True Positive + False negative)
- Specificity of a test is the ability to correctly detect who does not have a condition.
  - True Negative / (True Negative + False positive)

# Sensitive (and Specific) Love™

- Our favorite daft chemist, with the help of i, gets a huge grant to develop a test of happiness.
- She gets access to a population of 2000 people, half of whom are considered hypomanic (really, really, really happy).
- Can her test distinguish the two groups?





# Contingency Tables

- You can quickly figure out sensitivity, specificity and some other useful numbers by building a contingency table.
- Contingency tables show the number of people who are classified into mutually exclusive categories. You run into them whenever there are people classified as diseased or not diseased based on a cut point on a test.

# Examples of Sensitivity and Specificity

<i>group(Truth)</i>	<i>happiness(Guess)</i>		
	<i>Manic</i>	<i>Normal</i>	<i>Total</i>
<i>manic</i>	499	501	1000
<i>normal</i>	0	1000	1000
<i>Total</i>	499	1501	2000

Cut too high at 125

Sensitivity = .499

Specificity = 1

Cut too low 100

Sensitivity = 1

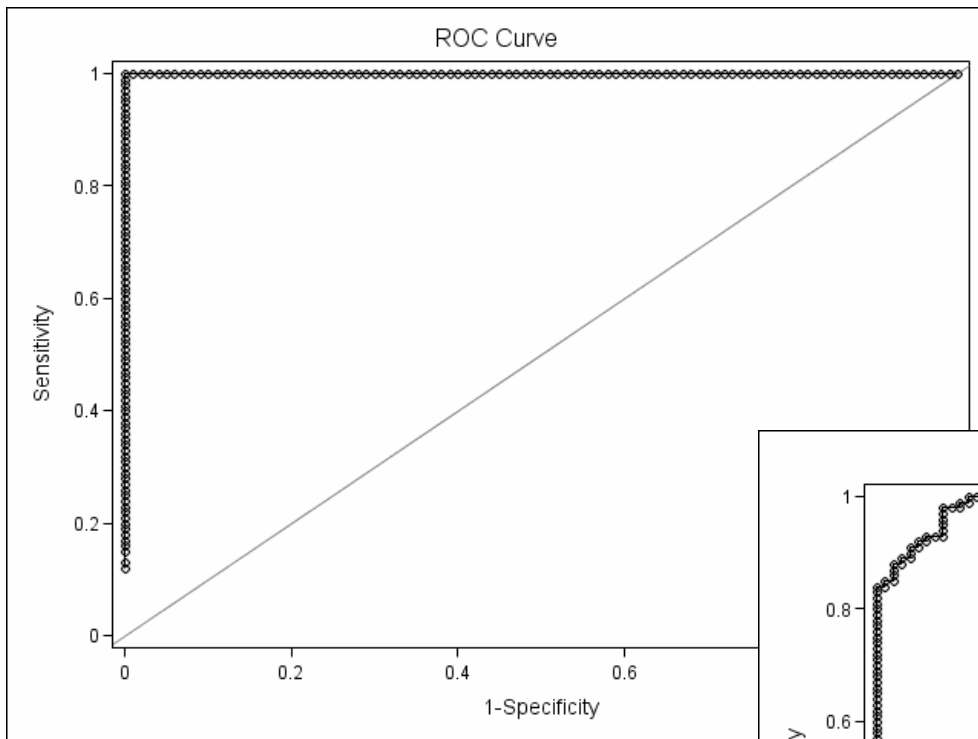
Specificity = .504

<i>group(Truth)</i>	<i>happiness(Guess)</i>		
	<i>Manic</i>	<i>Normal</i>	<i>Total</i>
<i>manic</i>	1000	0	1000
<i>normal</i>	496	504	1000
<i>Total</i>	1496	504	2000

# ROC Curves

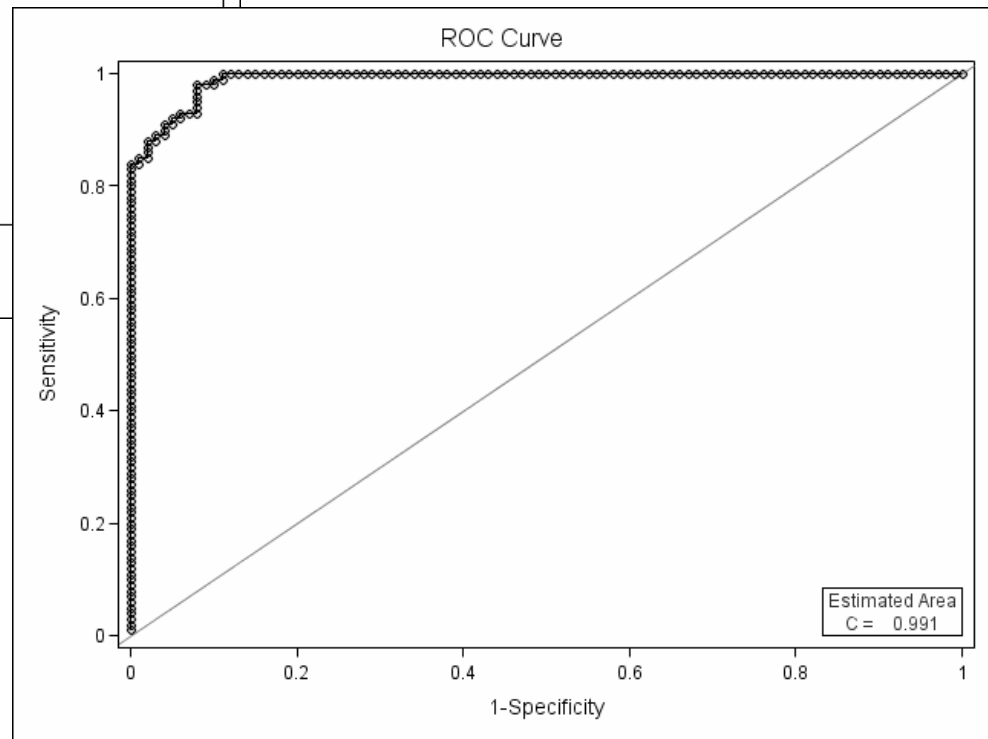
- Pretend you have a new test for mania with a range of scores from 1 to 100 (insanely blissful) and you have a pool of people who you know are or are not manic based on DSM criteria. You can say everyone is manic if they have a score of 1 and look at the sensitivity and specificity (100% sensitive, 0% specific). Then shift the criteria for calling people manic and keep repeating the process and plot the results. This information is plotted to make an ROC curve.

# ROC Example



← In an ideal case

More likely →



# Other Contingency Tables

- You will see contingency tables wherever you have a categorical (binary in this case) outcome and a categorical predictor.
  - Case control studies
    - Get a case and a control and look at previous exposure.
  - Prospective studies
    - Select people with and with exposure and see who gets diseased.
  - Cross sectional studies
    - Select people and see who does and does not have exposure and who does and does not have disease.

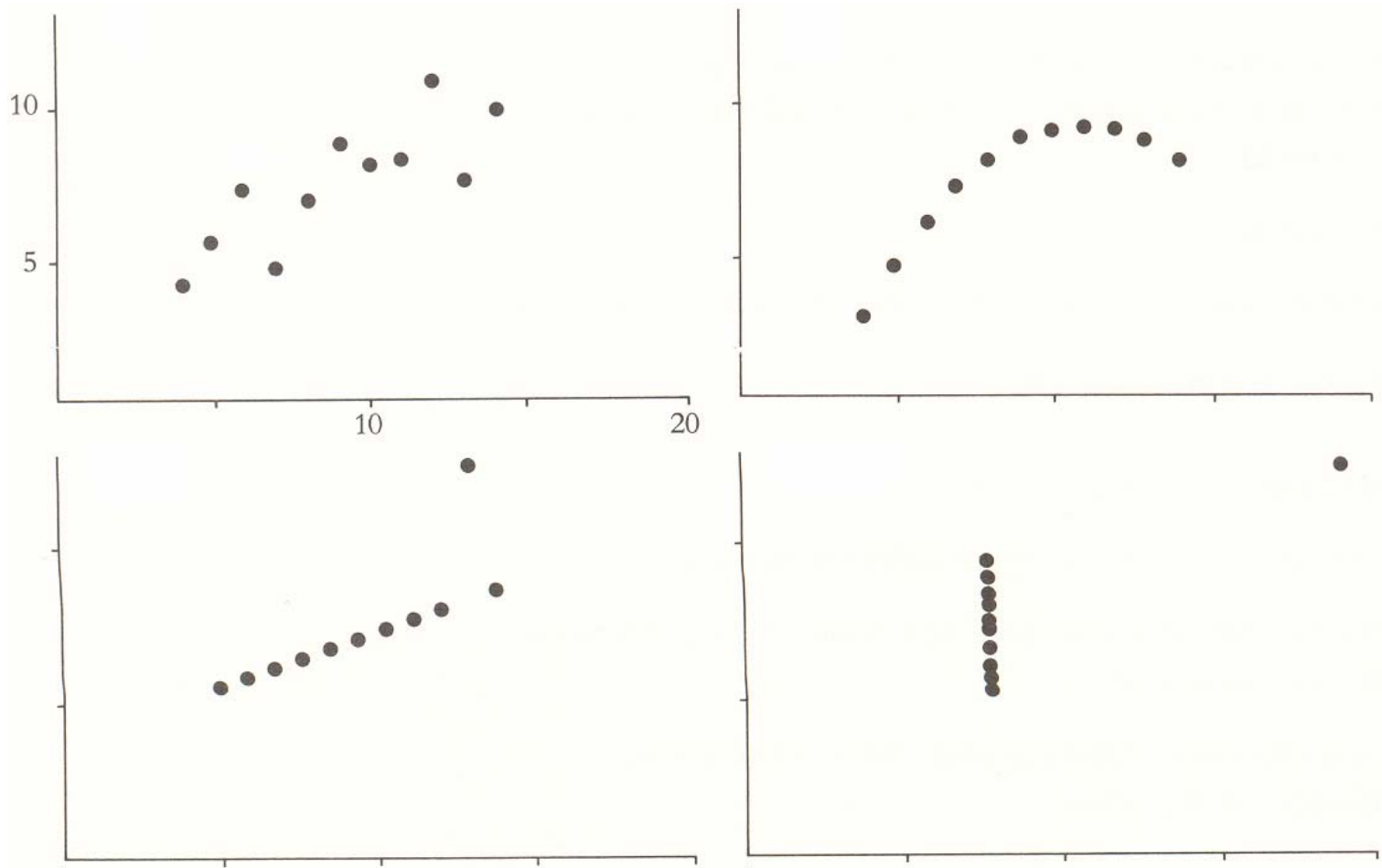
# Analyzing Contingency Tables

- When you have data in a contingency table you can work out an odds ratio.
- More on this in a bit..

# Correlation

- If a score on one measurement goes up (or down) systematically as another factor goes up and down, those two things are correlated.
- Conceptually, correlations are used when you are measuring two factors on the same person.
- There are two different common ways to measure correlation: Pearson's correlation coefficient and Spearman's.
  - Pearson's is used when you have two normally distributed variables.
  - Spearman's is used when the data is not normally distributed. It is a rank order statistic.

# Scatter Plot for Correlations



All have  $r^2 = .67$

# Correlation in Excel

- You can ask a simple program like Excel to generate the Pearson correlation but it doesn't give you all the information you will want.
  - Go to the tools menu
  - Pick “Add-ins...”
  - Check off Analysis ToolPack checkbox
  - Use the “Data analysis...” option on the tools menu.

# Correlation

- If:
  - your subjects are randomly sampled from a population
  - You have paired data on each person
  - You have independent observations
  - X and Y are measured independently
  - You are not controlling the values of the variables
  - X and Y are normally distributed
- Then you can square  $r$  ( $r^2$ ) and interpret it as the % of the variability of one variable that can be explained by the second variable. Call this the coefficient of determination.

# Correlation and Causation

- Do matches cause lung cancer?
- Does childhood ice-cream use cause adult heroin consumption?

# Confidence...

- The value you get for the correlation coefficient is not likely to be the true value in the population. If you measured only one or two people, you would not have much faith in your estimate but if you measured 10,000 people, you would probably trust your guess. You can quantify your guess by generating a confidence interval.

# What is a confidence interval?

- Suppose you were to get 100 samples from the population you wanted to measure and you know the true correlation coefficient. You could then work out the correlation coefficient in each sample and then put a range of +/- some amount for all guesses. You could tweak the range until 95% of the samples had the true population correlation coefficient covered by the range. The range (the estimate +/- some number) would be called the 95% confidence interval.

# Interpreting a Confidence Interval

- In real life you only can afford to get one sample but you can still ask a computer to work out a 95% confidence interval around your statistical estimate.
- The fundamental idea is that there is a true population value and you have a point estimate (the r-squared value in this case) from your sample with some “wiggle room” around it and using some math, 95% of the time the true population value is captured in the “wiggle interval.” So you can say you are 95% confident that your sample confidence interval has the true value.

# Confidence Limit Formulas

- If you use any reasonable software package, it will work out the confidence limits on practically any statistical estimate.
- The formulas differ depending on what you are estimating (a mean, a proportion, a correlation) but the interpretation will always be the same.
- You can work out different confidence intervals, for example, 90% confident you have the population value, 99% etc.
  - The larger the percentage, the wider the interval.
  - The larger your sample, the smaller the confidence interval.

# SD, CL, SEM oh my!

- You will frequently see graphics that have a mean plotted as a dot with whiskers extending up and down. These whiskers can be standard deviations, confidence limits on the point or standard errors.
- Conceptually, standard deviations describe your sample and confidence limits describe the precision of the estimate. Standard errors are used in determining the confidence limits.

$$SEM = \frac{SD}{\sqrt{N}}$$

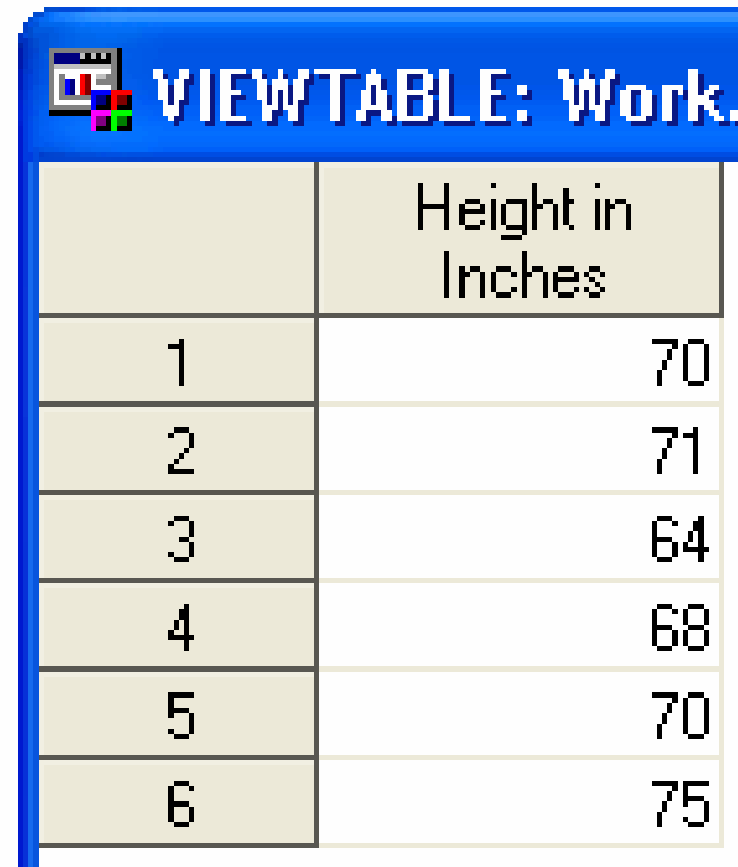
$$CLs = Mean \pm 2ish \times SEM$$

# Confidence limits at work...

- Our intrepid biochemist hires an evil lackey who we will call Igor. She tells Igor to slip a hefty dose of her concoction into the water cooler of the building next door.
- A week later she has Igor measure the subjects. One of his measurements is the (victims') "subjects'" height.

# What is Expected

- She guesses that if Igor got a typical sample of women at Stanford, the height will be 5 feet 5 inches tall. Igor skulks off and measures the height of the women. He observes 6 women at the tainted water-cooler and measures their height:



	Height in Inches
1	70
2	71
3	64
4	68
5	70
6	75

# Is the sample what you expected?

- What you expect to find is women that are 65 inches tall. It is far too easy to call this value the *expected value* so instead, statisticians will call it  $H_0$  or the *null hypothesis*.
  - That is pronounced “H zero”
  - Don’t pronounce it “hoe”
  - If you want to impress people, call it the null hypotheses.
- A null hypothesis is essentially what you expect if nothing “interesting” is going on in your study.

# Null Love™

- She decides that if her subject heights are very unusual, she will need to sack Igor. If the mean height is much greater (or smaller) than you would get from 95% of the samples of women who are 5'5", then she will decide that  $H_0$  was not correct and she will reject that hypothesis (and reject Igor). Instead she will accept the alternative that the heights are from a population of women that are taller or shorter than 5'6".

# Weird Love™

- If the population really is 5'5" you certainly could get a mean from a sample that is only 4'11" or 6'2". It would just be very weird.
- When you do experiments, you decide in advance just how weird the weirdness has to be before you reject the null hypothesis. This value is typically called a critical p value. Most people go with a p value of 5% or less.

# Bye Igor

- She asks Igor to analyze the data and to test to see if the mean value is significantly different from 65 inches. If the value would happen in less than 5% of the samples of size 6, she will reject the null hypothesis and conclude that the mean is different.

Statistics										
Variable	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err	Minimum	Maximum
theHeight	6	65.873	69.667	73.46	2.2564	3.6148	8.8657	1.4757	64	75

T-Tests			
Variable	DF	t Value	Pr >  t
theHeight	5	3.16	0.0250

# You can't prove it!

- In this case you conclude that the sample was different from the expected mean because the p-value was less than .05.
- P values do not tell you if something is clinically interesting. It just says if the results are weird (incompatible with the null hypothesis).
- They never *prove* that the null hypothesis is true or false. They can encourage you to accept or reject the null hypothesis.

# Multiple Comparisons

- You can decide to reject the null hypothesis if the data would be this weird 1 in 20 times ( $p < .05$ ) but then what happens if you do 20 tests? You are likely (64%) to reject the null just by chance.  
error =  $1 - (.95^{**20})$
- So when you do multiple comparisons, you need to demand much smaller p-values.

# What is hypothesis testing?

- Biostatistics is about describing data and testing to see if it is compatible with preconceived ideas.
  - Does eating garlic reduce blood pressure relative to a placebo?
  - Does taking OBC reduce risk of ovarian cancer in woman who have BRCA mutations?
  - Does taking a nap when working the night shift reduce driving errors at the end of the shift?
- Analyses start by saying nothing is going on, then testing to see if the data is compatible with this null hypothesis.

# Some Ways to Test Hypotheses

- Sample mean vs. a population mean
- Two sample means
- Two sample means paired
- One sample median vs. a population median
- Two samples rank ordered
- More than two sample means
- More than two sample means; each person measured repeatedly
- Do two groups of patients die at same rate?
- one sample t-test
- Two sample t-test
- Paired t-test
- Wilcoxon Signed-Rank test
- Wilcoxon Rank-Sum test
- ANOVA
- Repeated measure ANOVA
- Log Rank test

# How to Not Find a Significant Effect

- You will fail to reject the null hypothesis when there really is something important going on if:
  - You get a weird sample.
  - You are looking for tiny effects in a complex system.
  - You don't have enough people.

# Take Love™ ... be happy!

- Our chemist got 20 “volunteers” to take her concoction. I convinced her to randomly assign half of the people to a placebo. Her results were:

<i>drugged</i>	<i>happiness</i>		<i>Total</i>
	<i>happy</i>	<i>unhappy</i>	
<i>drug</i>	8	2	10
<i>noDrug</i>	5	5	10
<i>Total</i>	13	7	20

The odds of being happy were 4 times as likely among the drugged! This is a huge effect but the p-value is only .35

# More People Take Love™!

- With 10x more people...

<i>drugged</i>	<i>happiness</i>		<i>Total</i>
	<i>happy</i>	<i>unhappy</i>	
<i>drug</i>	80	20	100
<i>noDrug</i>	50	50	100
<i>Total</i>	130	70	200

The odds of being happy were 4 times as likely among the drugged!  
This is a huge effect with a p value of  $< .000001$

# What are the odds?

- Epidemiologists and statisticians and gamblers like odds.

$$\text{prob} = \text{odds} / (\text{odds} + 1) ;$$

$$\text{odds} = \text{prob} / (1 - \text{prob}) ;$$

Probability of an event	Odds of an event
0.10	0.11
0.20	0.25
0.25	0.33
0.30	0.43
0.40	0.67
0.50	1.00
0.60	1.50
0.70	2.33
0.75	3.00
0.80	4.00
0.90	9.00

# Power

- Your ability to detect an effect when it is present is called the power of a study. Before you do a study you can calculate power by looking at how variable your outcome is, how big the effect you expect to observe is and how many people you expect to have.
- Other things that impact power include:
  - How close your measure is to the construct you really want to measure
  - The ratio of people in different treatment groups.

# Power

- Procedures to calculate power are built into the major analysis packages (SAS has a procedure called proc power that I use).
- When you review medical literature and you find null results (they failed to reject the null hypothesis of nothing interesting going on) read or calculate what the power was. People typically aim for an 80% chance to see a statistically significant difference if there is a true difference.

# When Power Really Matters

	Is a real difference	Is no real difference
Reject Null	No Error (true positive)	Type 1 error $\alpha$
Fail to reject	Type 2 error $\beta$	No Error (true negative)

Low metastasis potential

	Is a really caner	No cancer
High PSA	No error (true positive)	False positive
Normal PSA	False negative	No error (true negative)

Highly aggressive breast cancer

	Is a really caner	No cancer
Positive image	No error (true positive)	False positive
Negative	False negative	No error (true negative)

# Experimental Design

- Our researcher wants to test the impact of Love™. She does a series of experiments to determine the toxicity of Love™ and to identify side effects and determines how long Love™ maintains its impact. She is ready to do large scale testing on “healthy, normal” people. She talks to people at SPCTRM about how to design the experiment and it is suggested that she conduct a triple blinded experiment.

# Conducting Experiments

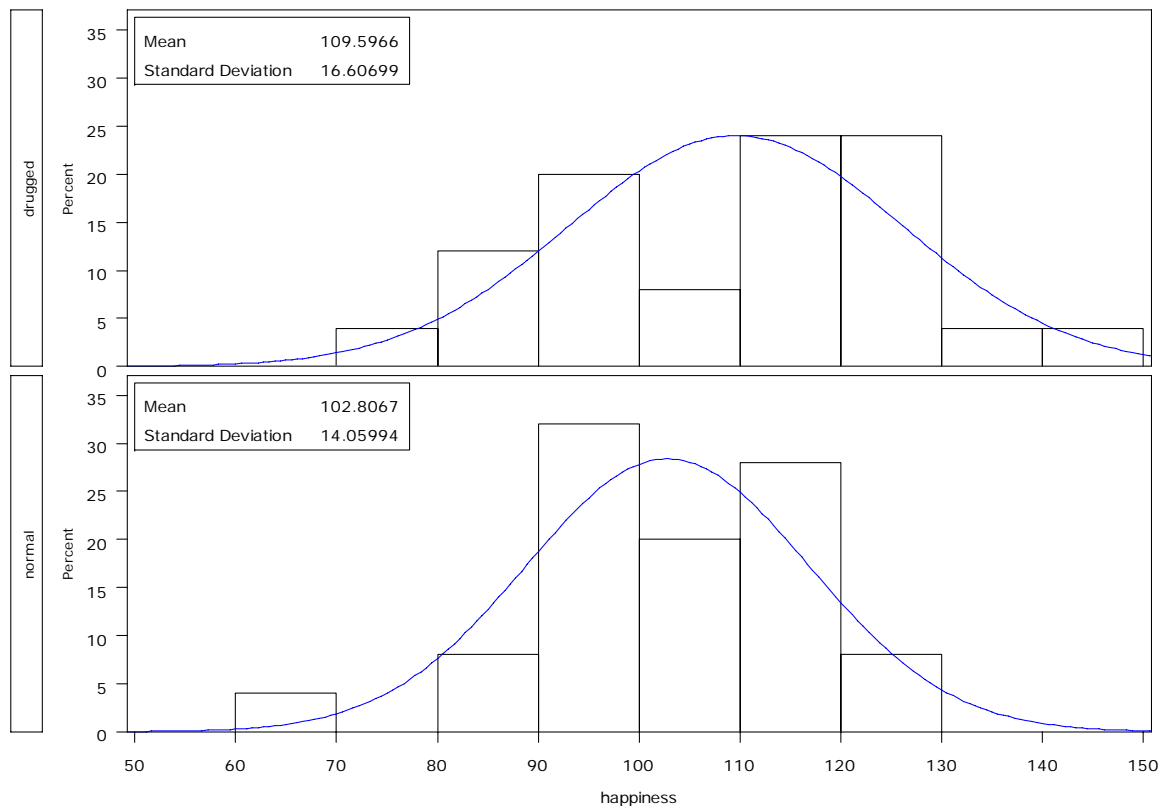
- Take a random (or representative) sample of the population you wish to generalize to.
- Randomly assign people to treatment groups. In theory, this balances out all extraneous factors other than treatment.
  - 2 of every 4 people are assigned to the drug – blocked randomization.
- Triple blind the study.
  - subjects don't know what treatment they are taking
  - the staff that interact with the subjects don't know what treatment they are taking
  - the people doing the analysis don't know what treatment subjects are on while they are doing the analysis
- Failing to blind causes biases.
  - Systematic errors

# Impact of Love™ on Normal, Healthy Adults...

- Preliminary analyses of Love™ indicate that it is perfectly safe when taken in small doses. What is the overall impact of Love™?
  - 100 people are given Love™ and their happiness is measured.

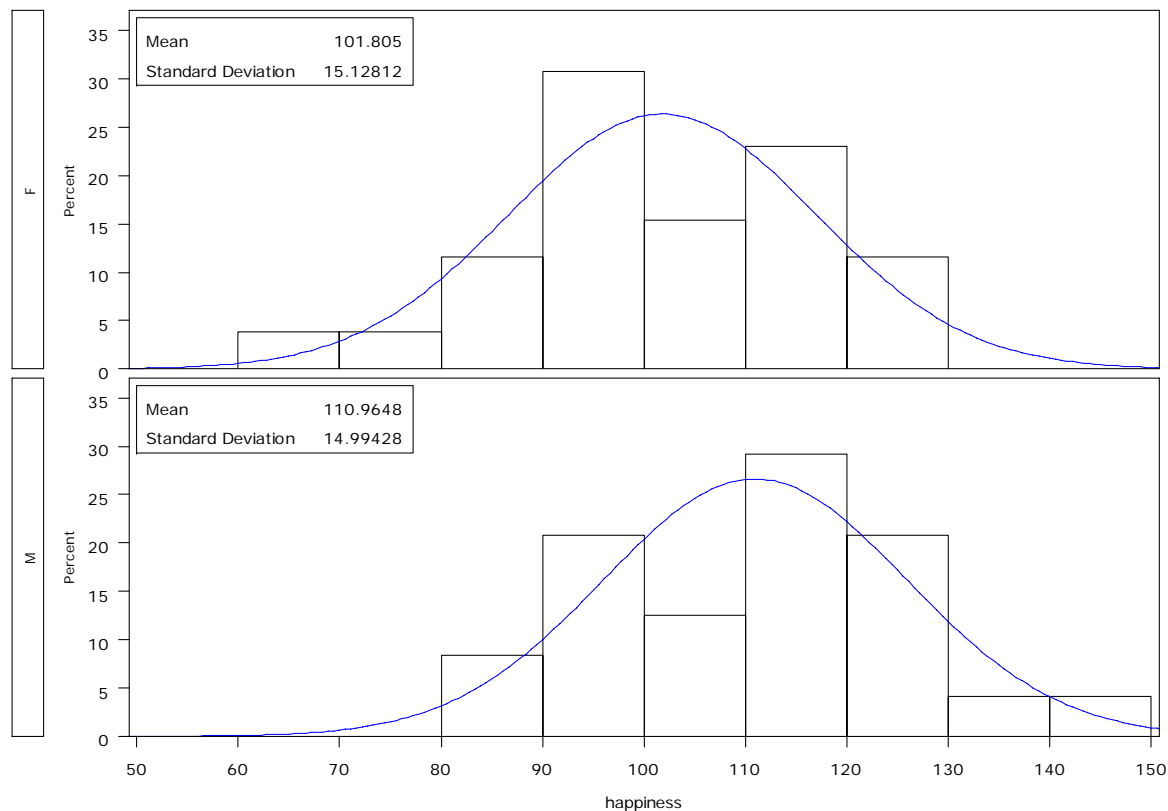
# Love™ Makes no Difference

- There is a lot of variability and very little difference in the means  $p < 0.1253$ .

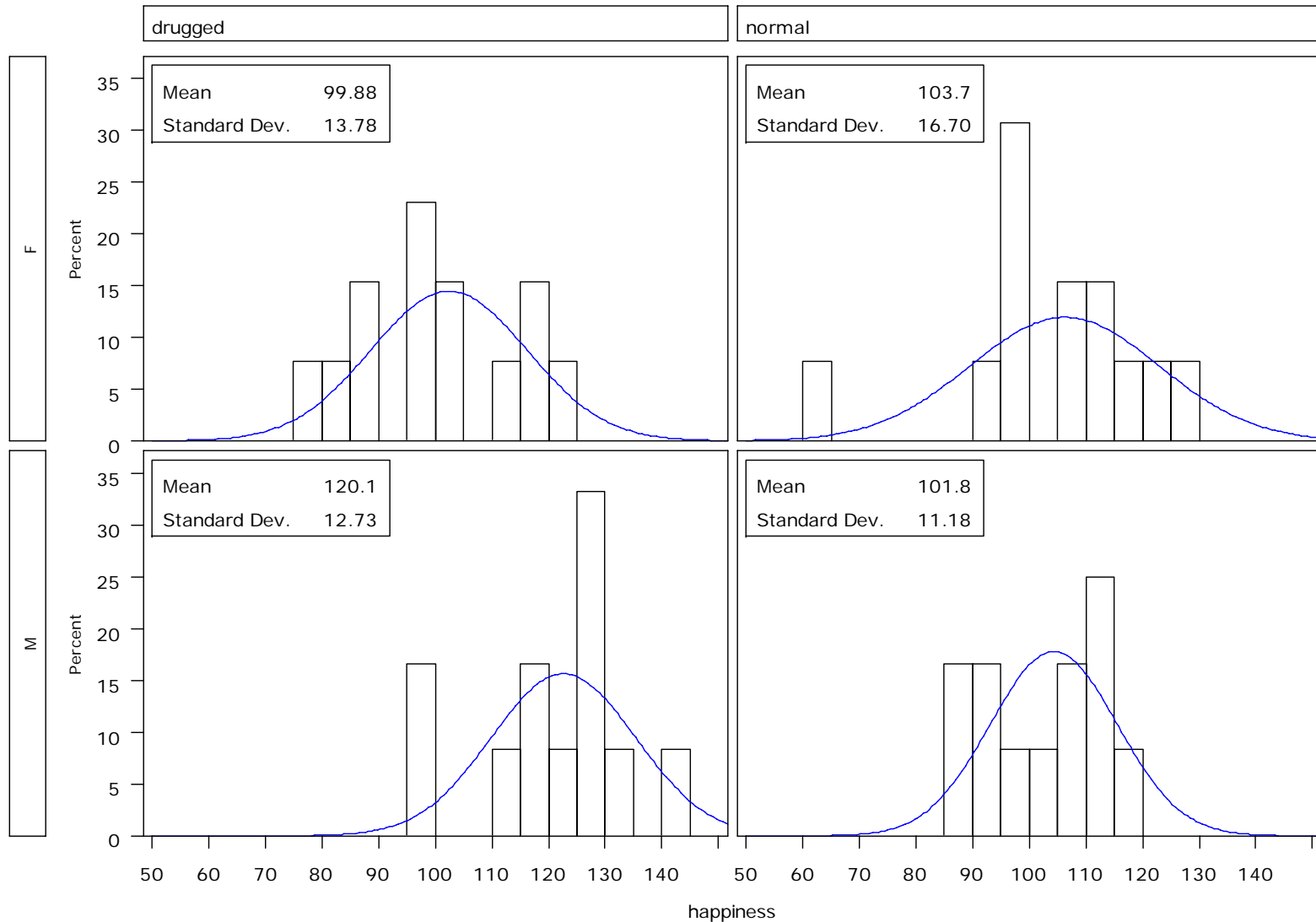


# Who is happier?

- The men are happier on average and it looks statistically significant  $p < 0.0368$ .



# Gender and Love Interacting....



# Love<sup>TM</sup> is complicated.

- In this case the main effect of the drug is hidden away in the interaction between gender and the drug. Whenever there are interactions, you need to be very careful in interpretation.
  - Graph everything!
- You could predict happiness with an equation that had baseline happiness + an impact of being male + impact of being on the drug.

# Regression Lines

- The line that was fit through the data is called an ordinary least squares regression line.
- With ordinary least squares you anchor the regression line at the means of the  $x$  and  $Y$  values and then jiggle the line around until it is as close as possible (measuring up and down) to all the plotted values.

# Predicting the Future

- Regression is useful when you want to make predictions. Ordinary least squares is useful when you want to make predictions about a variable that is approximately normally distributed. You feed your sample data into a statistical program and then ask it to give you a formula that can be used with future data.

# Types of Regression

- Ordinary least squares regression predicts a normally distributed outcome.
- Logistic regression predicts a binary (or categorical outcome).
- Cox regression predicts how long a person will live (or really how long until they will last until an event).
- Nonparametric regression

# Assumptions for Regression

- What type of regression you are doing requires you to make different assumptions about the data. For a very readable discussion, see Motulsky's *Intuitive Biostatistics*.
  - Random sample of people
  - Independent samples

# Do-it-yourself Biostatistics

- There are many software packages for doing statistics and data management.
- Use whatever package you can get help on.
  - SPSS: superb for usability
  - S-plus: superior graphics, decent user interface, hard language behind the scenes
  - SAS Enterprise Guide: good user interface and powerful
  - SAS: only if you have help; extremely powerful
  - STATA: less useful but easier to learn than SAS
  - R: arguably the most user-hostile program ever written

# Tools for Examining Data Graphically

- Excel – ubiquitous with horrible defaults
- Delta Graph – very solid graphics
- Paintshop Pro – great for putting finishing touches on graphics
- GraphPad – nice introductory level analysis and graphics toy
- Tableau – fantastic new package for exploratory data analysis

# My Favorite Intro Books

- Books to use to learn basic biostatistics
  - Intuitive Biostatistics by Motulsky
  - Biostatistics: The Bare Essentials by Norman & Streiner
  - Common Statistical Methods for Clinical Research with SAS Examples by Walker

[www.stanford.edu/class/hrp223/2002f/books.html](http://www.stanford.edu/class/hrp223/2002f/books.html)

# My Favorite Classes on Campus

- HRP 223: Data Management and Statistical Programming
  - Applied class on how to manage data and do statistics
- HRP 259: Introduction to Probability and Statistics by Kobb
  - From theory to application of statistics
- HRP 261: Discrete Data Analysis by Kobb
  - How to deal with categorical outcomes

# Resources on Campus

- Statistical Software Support Group  
[library.stanford.edu/services/social\\_sci\\_data\\_soft/](http://library.stanford.edu/services/social_sci_data_soft/)
- Statistical department help  
[www-stat.stanford.edu/consulting/index.html](http://www-stat.stanford.edu/consulting/index.html)
- SPCTRM  
<http://clinicaltrials.stanford.edu/>