Silly or Pointless Things People Do When Analyzing Data: 1. Conducting a Test of Normality as a Precursor to a *t*-test

Bruce Weaver

Northern Health Research Conference

June 10-11, 2011

Speaker Acceptance & Disclosure

- **I have no affiliations, sponsorships, honoraria, monetary support or conflict of interest from any commercial source.**

- ***However***…it is only fair to caution you that this talk has ***not*** undergone ***ethical review*** of any sort.

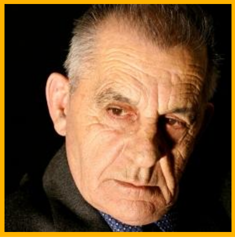
- Therefore, you listen ***at your own peril***.

On Sermons and Stats Lectures

"The secret of a good sermon is to have a good beginning and a good ending; and to have the two as close together as possible."

— George Burns

- This probably applies to stats lectures too, so I'll make *every* effort to keep it snappy



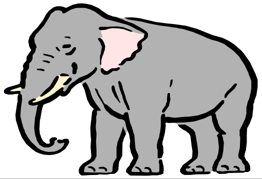
Now to the serious stuff…

Statistics textbooks often list the following assumptions for the unpaired *t*-test, **usually in this order**:

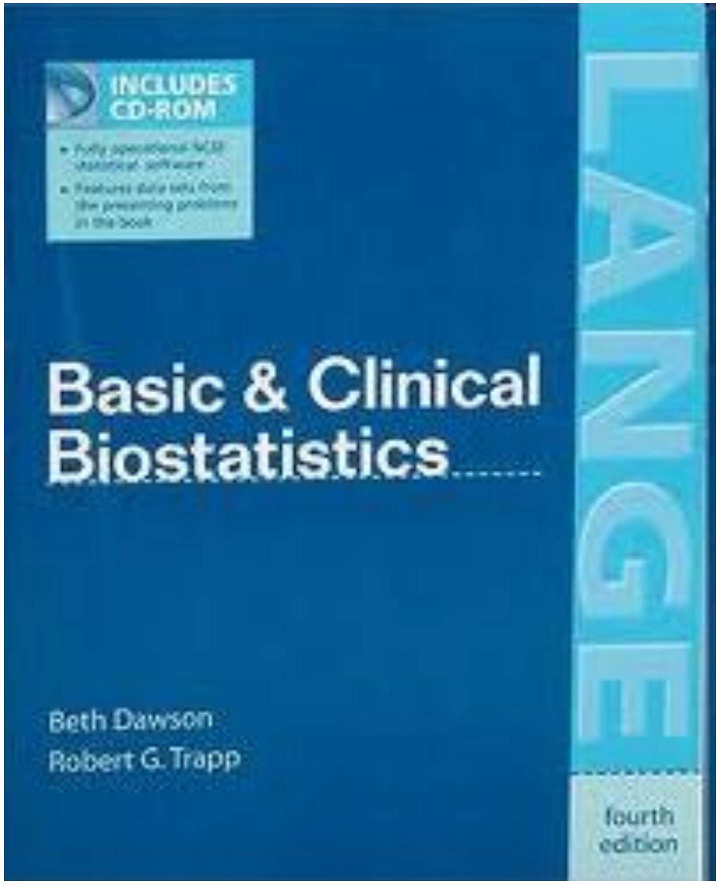
- The populations from which the two samples are (randomly) drawn must be

  1) **normally distributed** with
  2) **equal variances**, and
  3) each observation must be **independent** of all others.

Yes, that looks familiar.

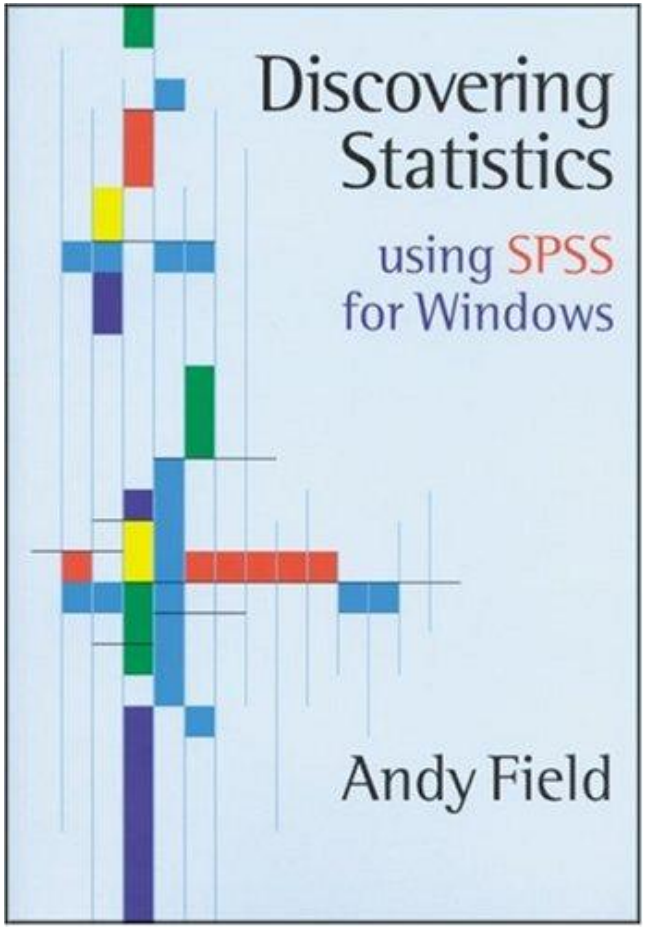


Example from a Biostatistics Textbook



1. "…the observations in *each group* follow a normal distribution."
2. "The standard deviations (or variances) in the two samples are assumed to be equal"
3. "…independence, meaning that knowing the values of the observations in one group tells us nothing about the observations in the other group"

Example from an Applied Stats Textbook: Assumptions for Parametric Tests



First Edition (2000)

1. **Normally distributed data**: It is assumed that the data are from a normally distributed population.
2. **Homogeneity of variance**: … the variance should not change systematically throughout the data.
3. **Interval data**: Data should be measured at least at the interval level.
4. **Independence**: …behaviour of one participant does not influence behaviour of another.

Normality Assumption Often Listed First

- The assumption of sampling from normally distributed populations often appears ***first*** in the list
- This can lead users of statistics to conclude that normality is the most important assumption
- It is **not** – the ***independence*** assumption is by far the most important one…but I don't have time to talk about that today

How unfortunate! It sounds like a real *humdinger* of a topic.

Some Books Recommend Testing for Normality Prior to Running a t-test

- E.g., Field (2002) says to run a test of normality on the dependent variable
- In an example, he runs the **Kolmogorov-Smirnov** test of normality (with Lilliefors correction), and finds that it is ***statistically significant***
- What does that mean?

Who cares?! Did you say ***Smirnov***? Top me up please!

Interpreting Tests of Normality

- For any test of normality:

  H_0: Sample is drawn from a **normal** population

  H_1: Sample is drawn from a **non-normal** population

- If test of normality is statistically significant ($p \leq .05$), you conclude that the sample is from a ***non-normal*** population

- If test of normality is not statistically significant ($p > .05$), you have *insufficient evidence* to reject the null hypothesis—so you proceed as if the population is normal

This is what Field (2000) found

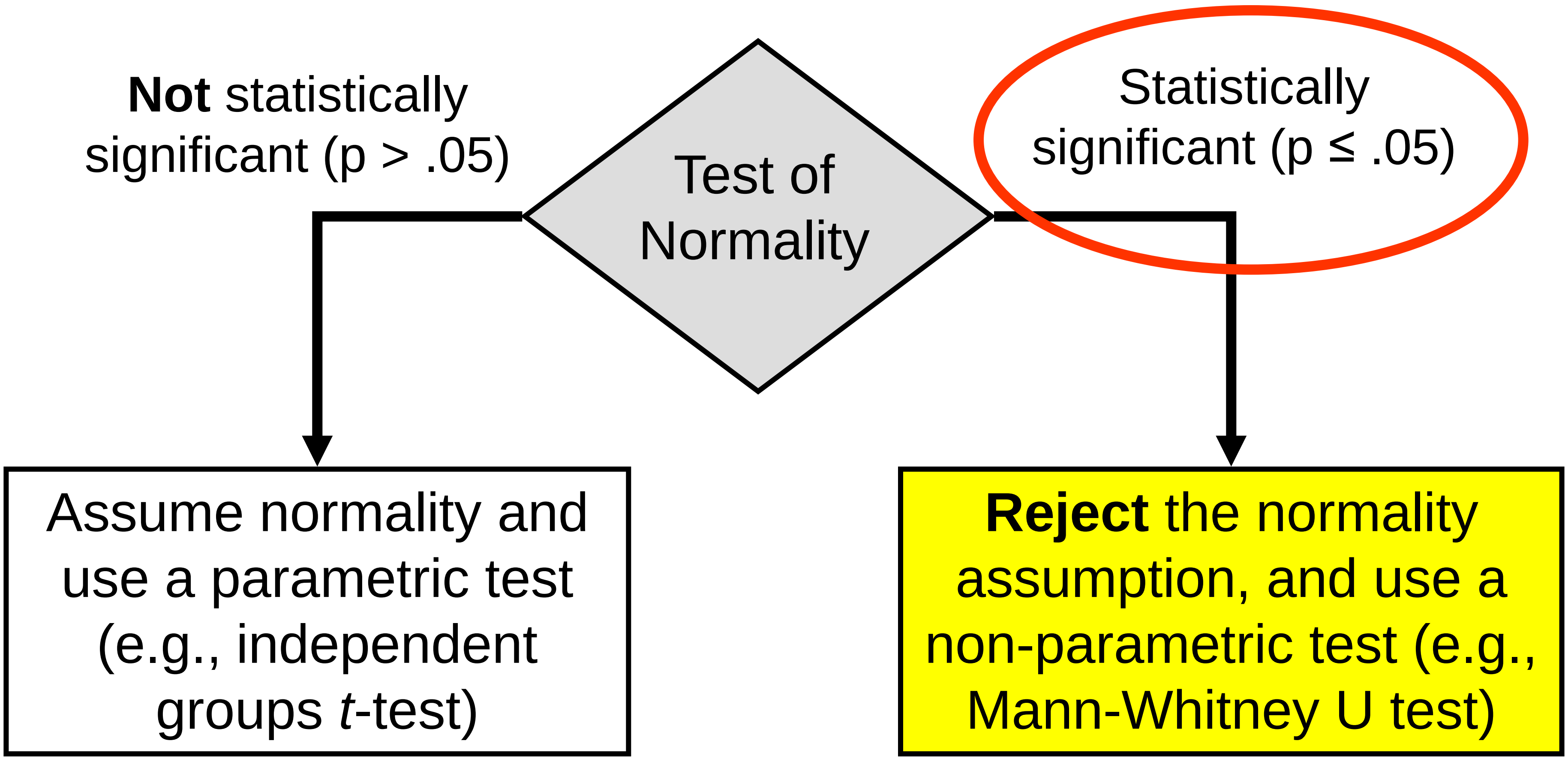
Okay, so what now?

Andy Field

"... we cannot use a parametric test, because the assumption of normality is not tenable." (Field 2000, pp. 48-49)

He then recommended using the **Mann-Whitney U test** (a rank-based test) instead of the *t*-test

Recap of Field's Procedure



Now ***ho-o-old*** on there, Baba-Louie!

Huh?

Quick-Draw McGraw

Baba-Louie

Let’s not forget what George Box said about normality!

Quick-Draw McGraw

Okay...

Baba-Louie

No ***real*** data are normally distributed

George Box

"…the statistician knows…that **in nature there never was a normal distribution**, there never was a straight line, yet with normal and linear assumptions, **known to be false**, he can often derive results which match, to **a useful approximation**, those found in the real world." (JASA, 1976, Vol. 71, 791-799; emphasis added)

Famous statistician and textbook author—and son-in-law of Sir Ronald F. Fisher

So the populations are **never** truly normal, at least not if you're working with real data.

Why is normality listed as one of the assumptions then?

What the textbooks ***should*** say

Yours truly

If one was ***able*** to sample randomly from two *normally distributed* populations with *exactly* equal variances (and with each score being independent of all others), then the unpaired *t*-test would be an **exact** test.

Otherwise, it's an **approximate** test.

Obscure statistical curmudgeon from NW Ontario—no relation to R.F. Fisher.

Approximate.
Is that bad?

No, not
necessarily.

Another Great Comment from Box

George Box

All models are wrong.
Some are useful.

i.e., they are approximations!

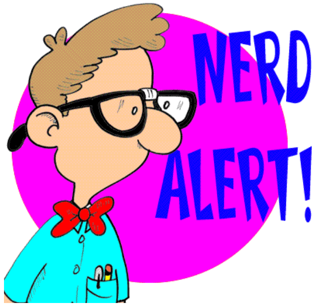
A ***New*** Question: Is it ***Useful***?

- From this point of view, the important question is not whether the populations we've sampled from are normal – **we know they are not**

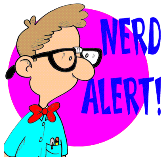
- Rather, the important question is whether the approximation is good enough to be ***useful***

Under what conditions is the approximation good enough to be useful?

To answer that, we need to look more carefully at how z- and t-tests really work.



How z- and t-tests **really** work

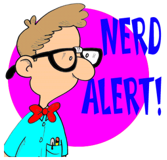


Common format for all z- and t-tests:

$$z \text{ or } t = \frac{\text{statistic - parameter}|H_0}{SE_{statistic}}$$

- Numerator = a statistic minus the value of the corresponding parameter under a true null hypothesis
- Denominator = the standard error of the statistic in the numerator

Example 1: Single-sample t-test



$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$$

Statistic = sample mean

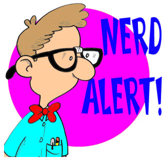
Parameter under a true H_0 = pop. mean

SE of the mean

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}}$$

($df = n$ - 1)

Example 2: Unpaired t-test



Statistic

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}}$$

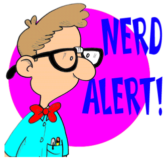
Parameter under True H_0

SE of the statistic in the numerator

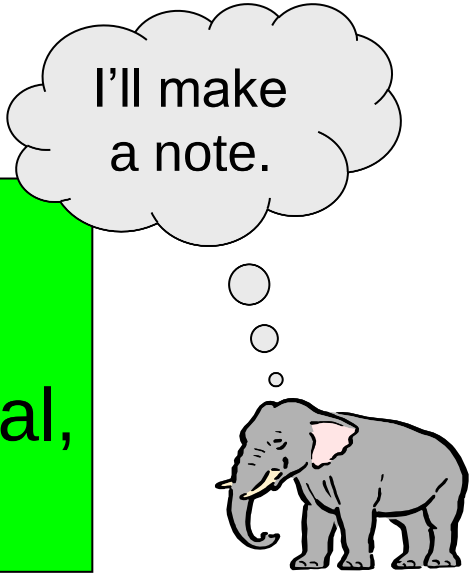
$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\left(\frac{s^2_{pooled}}{n_1} + \frac{s^2_{pooled}}{n_2} \right)}$$

($df = n_1 + n_2$ - 2)

The **KEY** Point

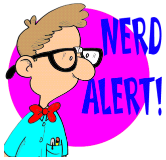


$$z \text{ or } t = \frac{\text{statistic} - \text{parameter}|H_0}{SE_{statistic}}$$



- It is the **sampling distribution** of the **statistic** in the numerator that must be normal, or at least approximately normal, in order to have a good test

The Central Limit Theorem (CLT)

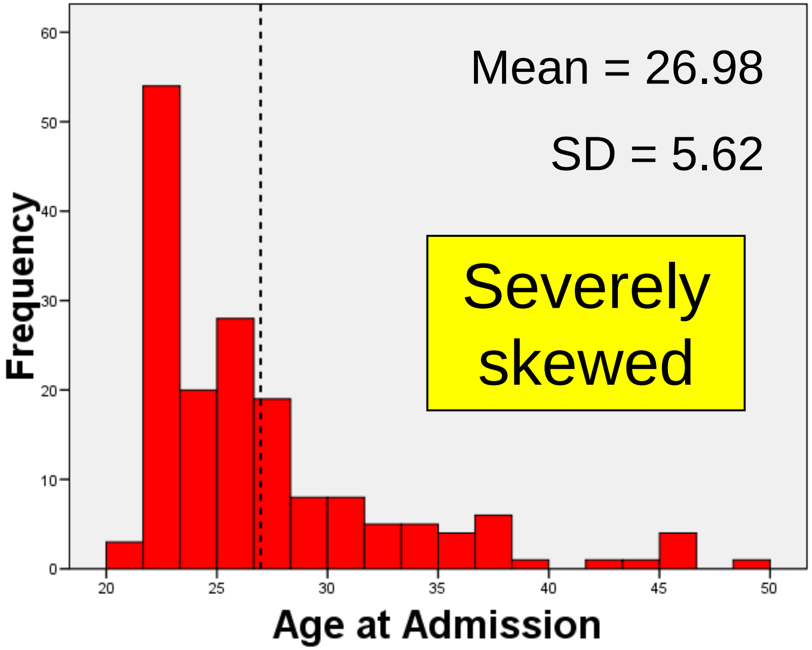


$$z \text{ or } t = \frac{\boxed{\text{statistic}} - \text{parameter}|H_0}{SE_{statistic}}$$

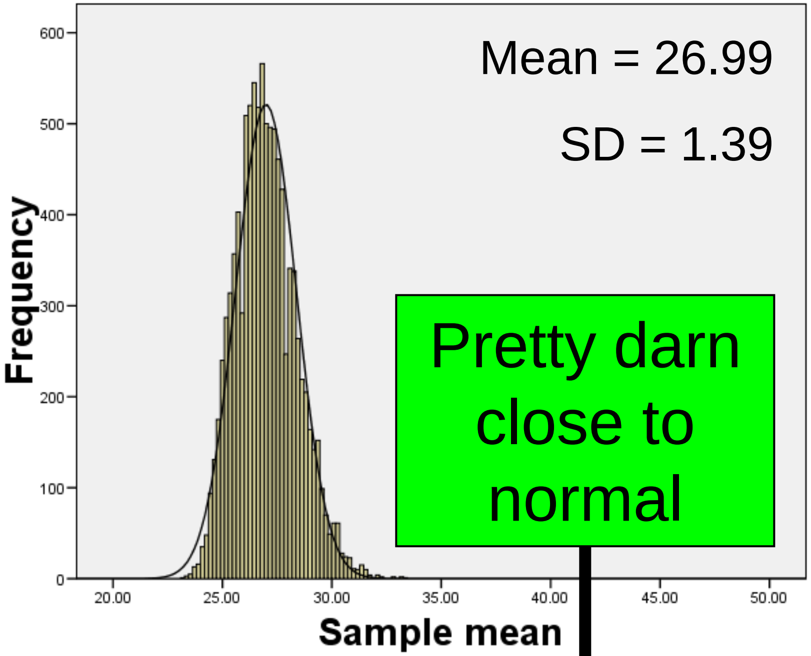
- The CLT tells us that as the sample size **increases**, the sampling distribution of the statistic **converges** on a **normal distribution**, regardless of the shape of the raw score distributions
- And *n* does not have to be all that large—see example on next slide with $n = 16$

An Example with 10,000 Samples of n = 16 from a Skewed Population

The Population Distribution



Distribution of **Sample Means** (for samples of n = 16)



Close enough for the Normal distribution to be a ***useful*** approximation

So in other words, the **larger** the sample size, the **less important** normality of the population distribution is, right?

This is ***correct***, Baba-Louie!

But at the same time, as the sample size increases, tests of normality become more and more ***powerful***.

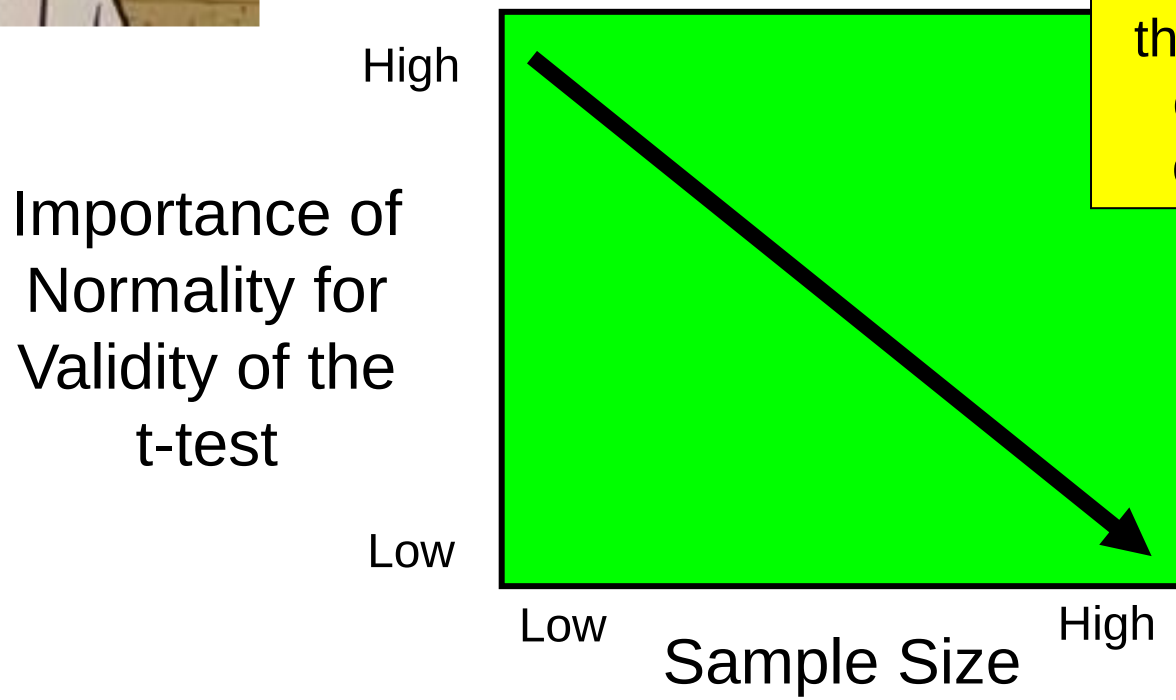
That's good, isn't it Quick-Draw? ***The more POWER, the better***! Right?

Tim "the *Stats*-Man" Taylor & Al

Not really so in this case, Tim.

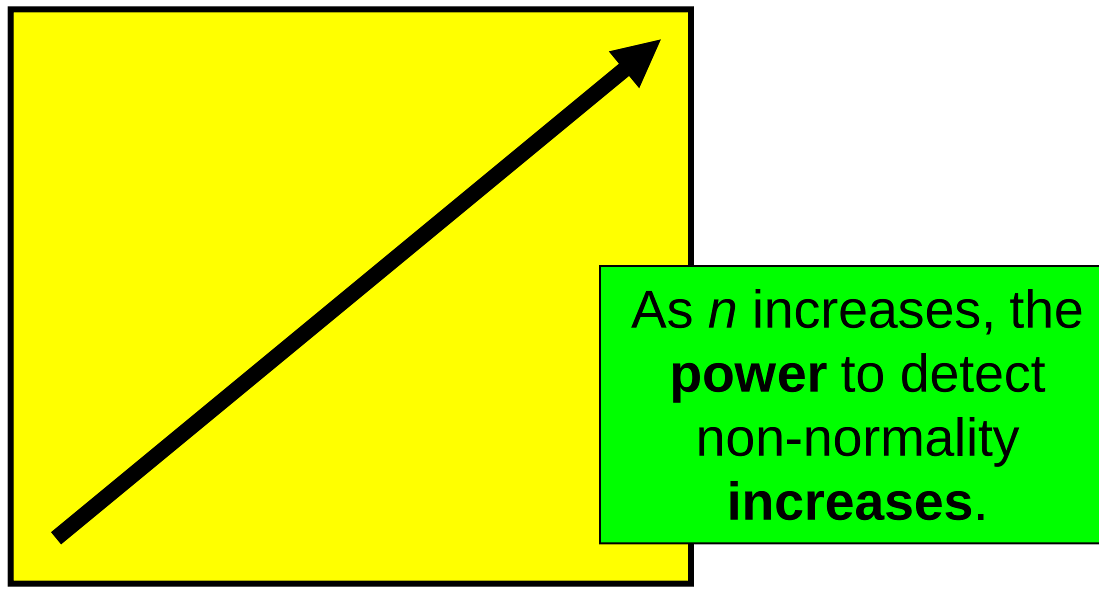
Pay attention now, Tim. Quick-Draw is about to make a very good point.

As *n* increases, **normality becomes less and less important** for the validity of the *t*-test.

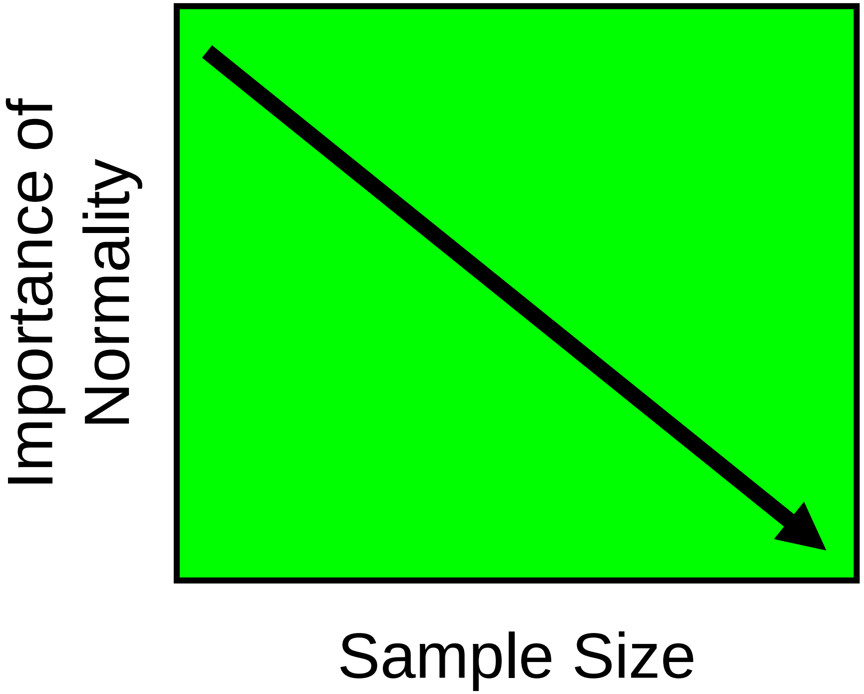


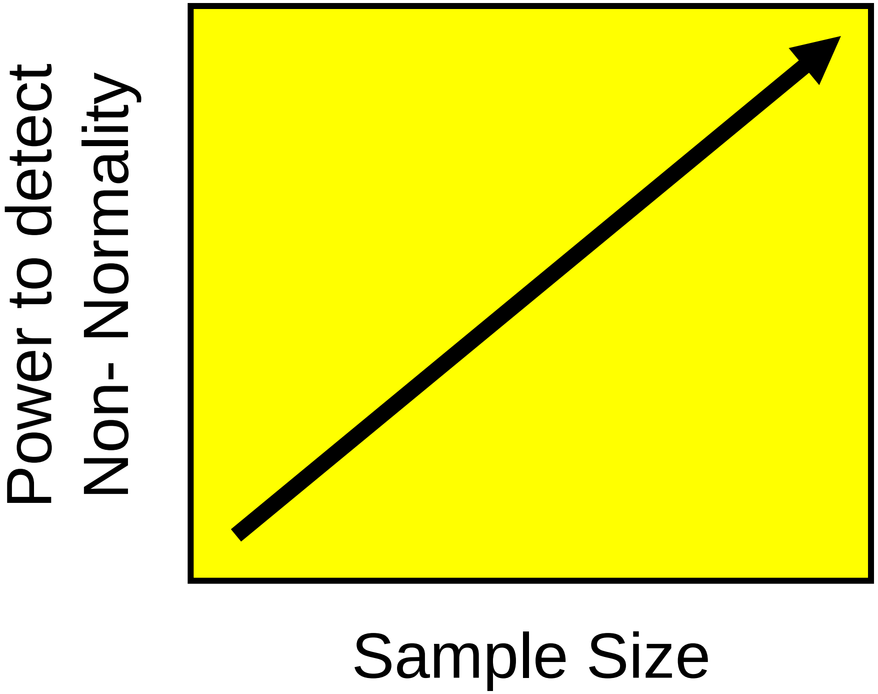
As *n* increases, the importance of normality **decreases**.

But at the same time, tests of normality become **more and more likely to detect significant non-normality**.



Recap of Quick-Draw's Point





- As n increases:
  - The importance of normality **decreases**
  - The power to detect non-normality **increases**

At cross-purposes!

And ***that***, my friends, is why tests of normality are really quite ***useless*** as precursors to t-tests or other parametric tests.

Yes, I see what you mean, *Queeks-Draw*.

Robustness of the t-test to Non-Normality of the Populations

Some Examples—*Time Permitting*

Click Here to Continue the Presentation

Click Here to Skip to the Summary

Robustness of the t-test to Non-normality

- The upcoming figure shows performance of the t-test when sampling from populations of various ***non-normal*** shapes

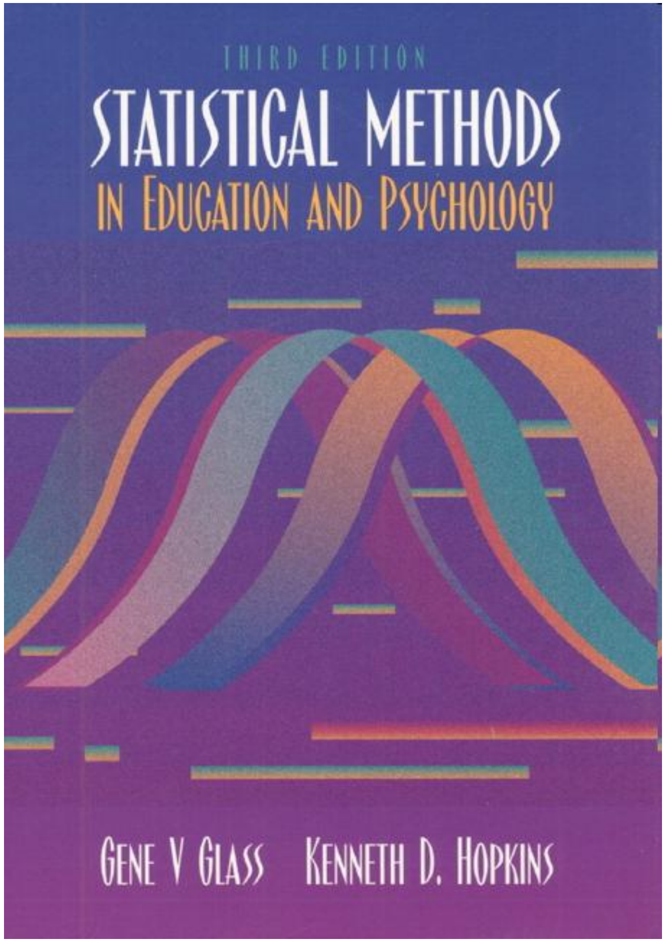
- Performance is measured by how closely **the actual proportion of Type I errors** matches the pre-determined alpha level – e.g., if you set alpha to .05, the actual proportion of Type I errors should be close to .05 for a good test

Cochran's Criterion for Acceptable Test Performance

- Cochran (1942) suggested allowing a ***20% error in the actual Type I error rate***—e.g., for nominal alpha = **.05**, an actual Type I error rate between **.04 and .06** is acceptable

- Cochran's criterion is admittedly arbitrary, but other authors have generally followed it (or a similar criterion) – so we will apply it here

Thanks to Gene Glass for Providing the Upcoming Figure



Gene V. Glass

Figure 12.2 from Glass & Hopkins, *Statistical Methods in Education and Psychology* , 3rd Edition

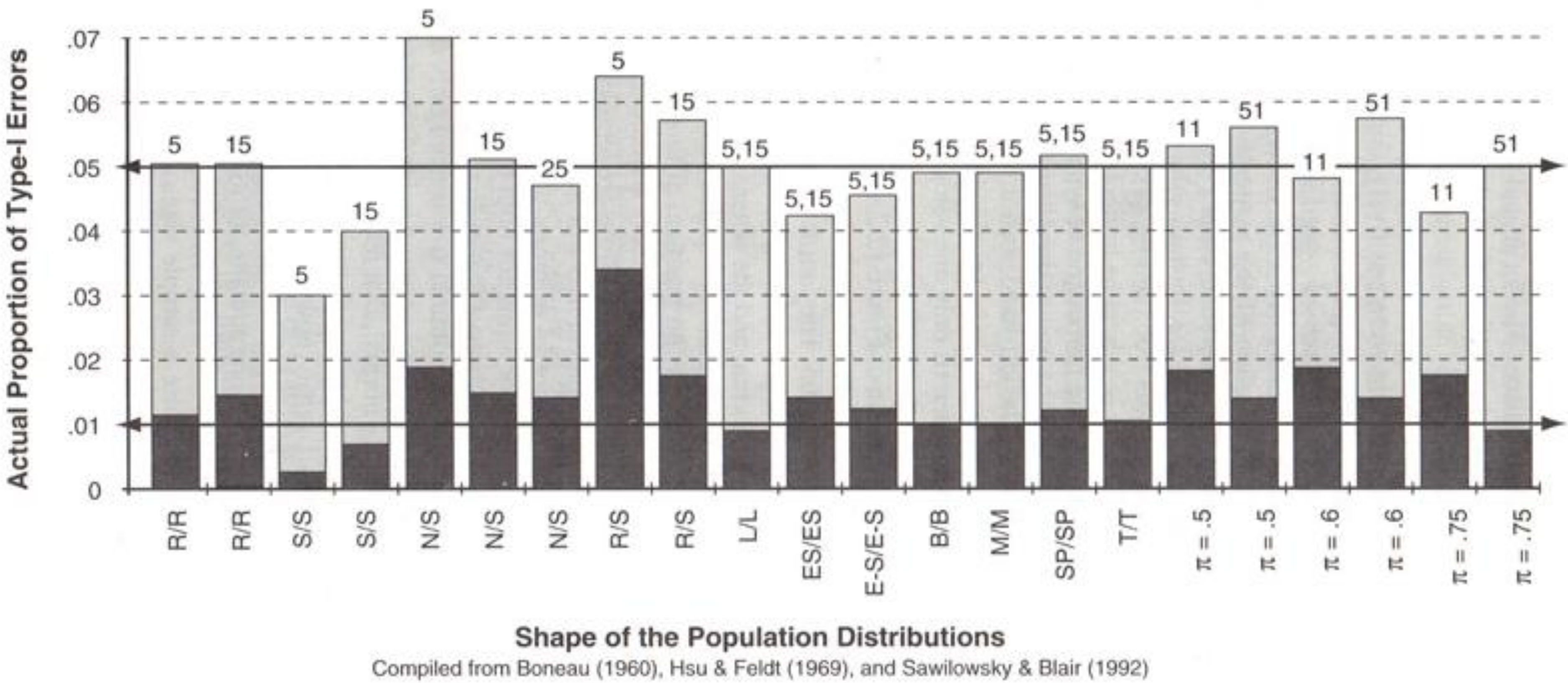
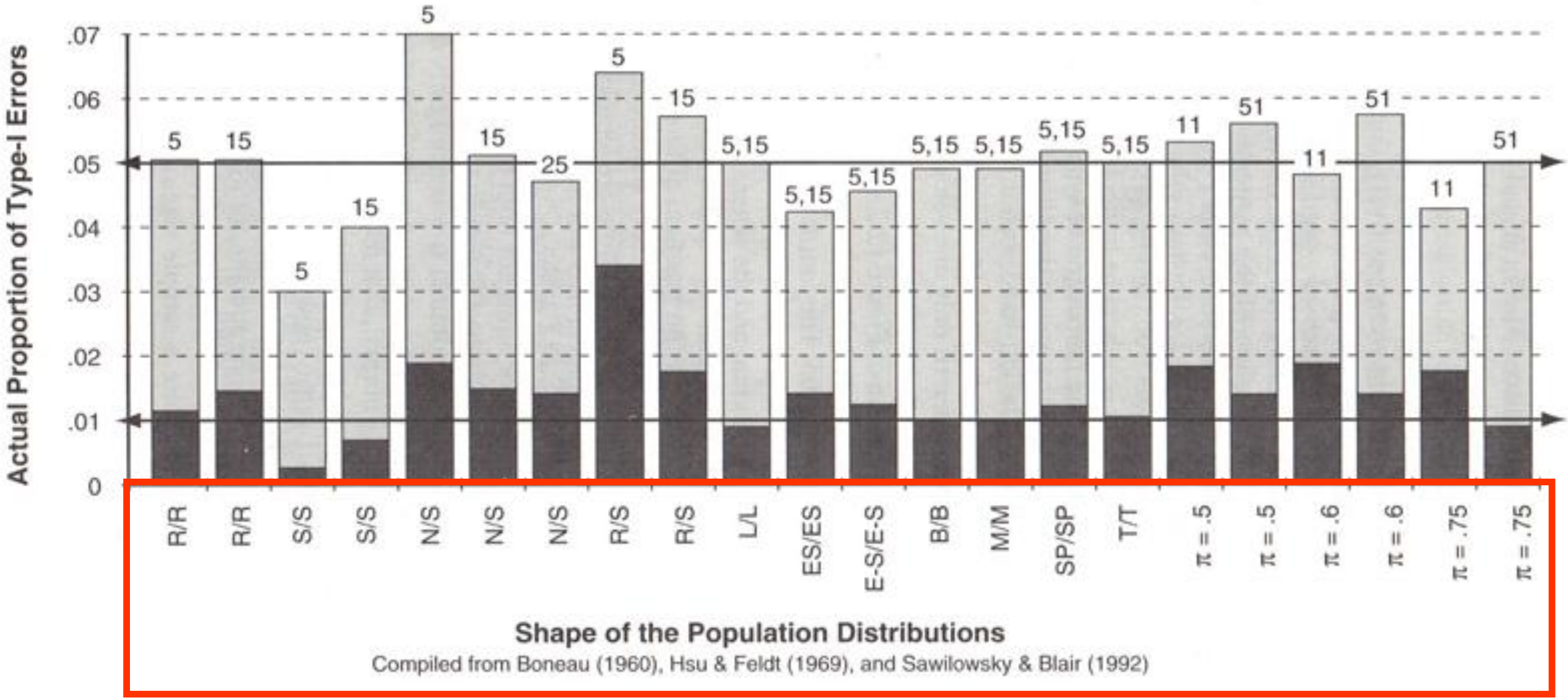
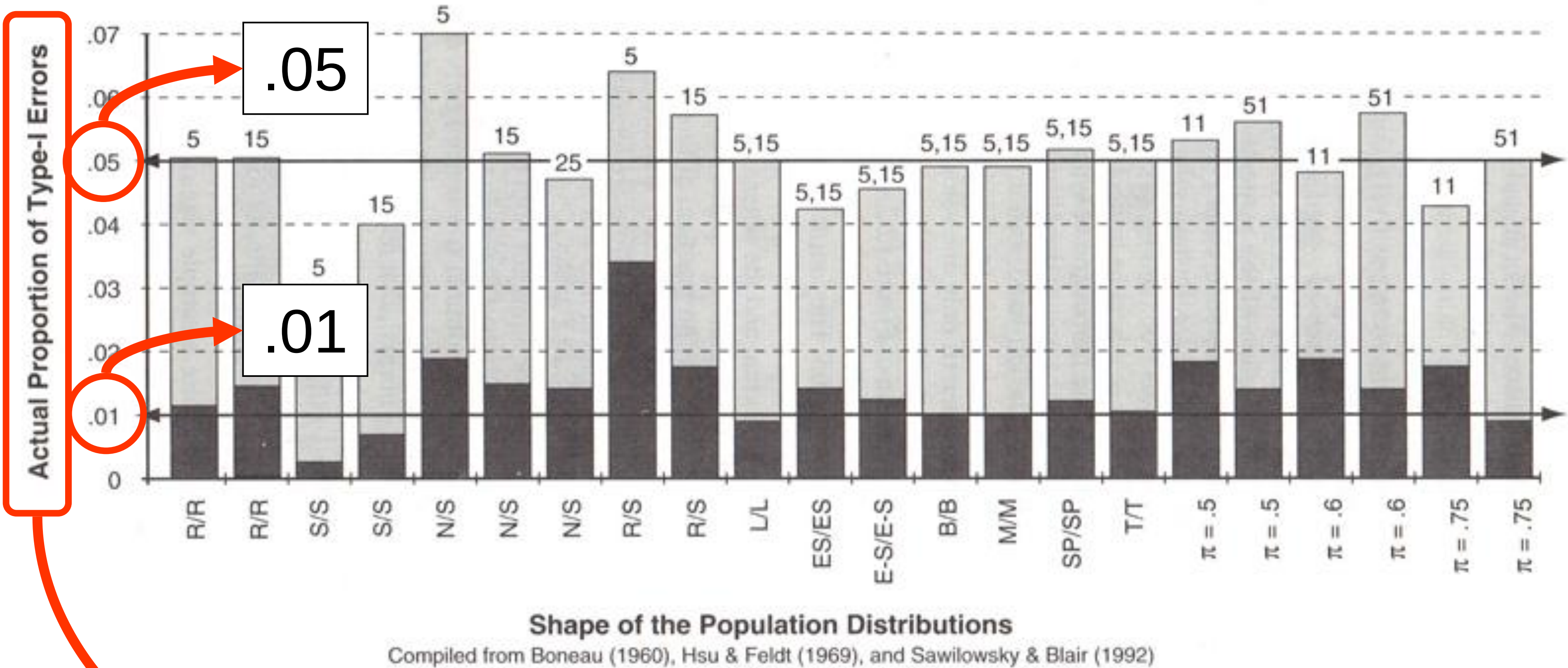


Figure 12.2 from Glass & Hopkins, *Statistical Methods in Education and Psychology* , 3rd Edition



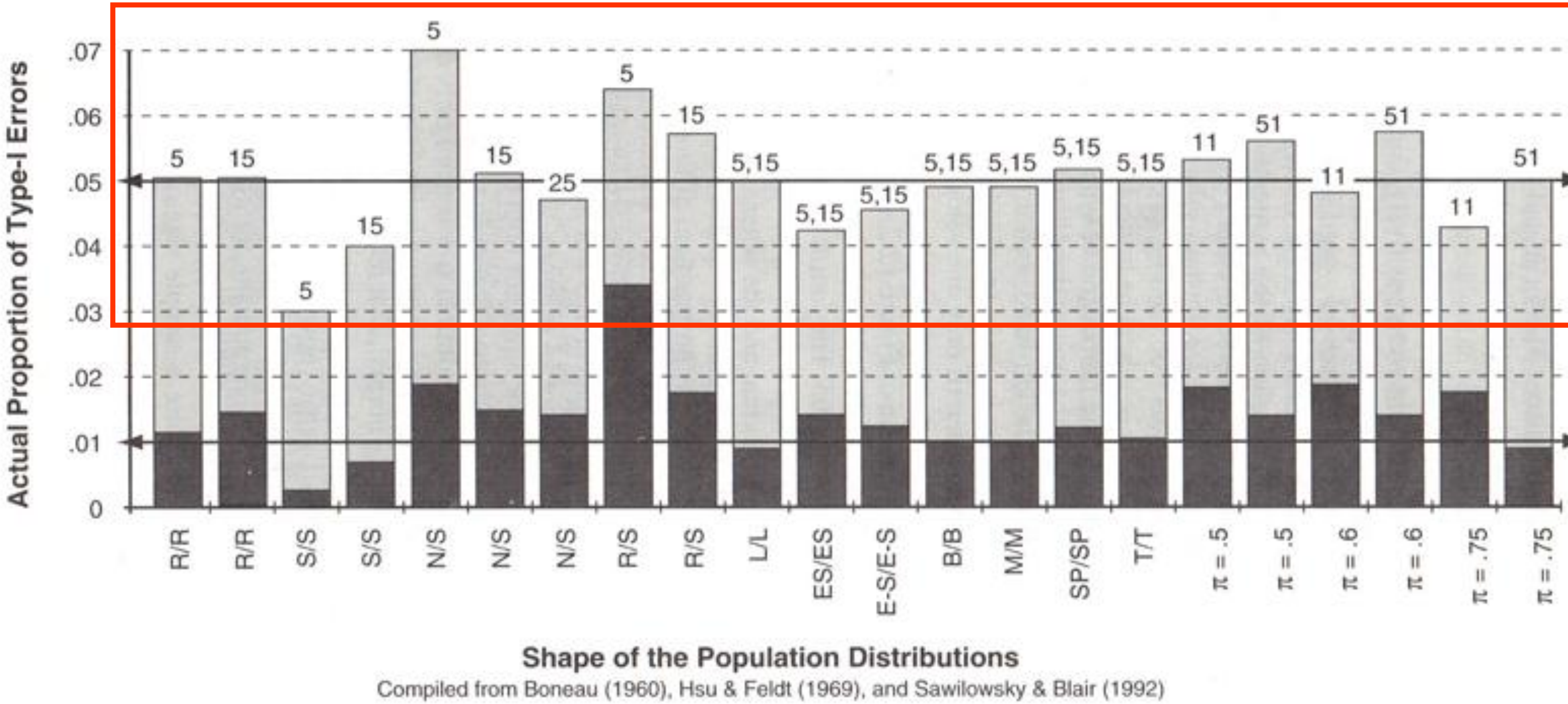
R = rectangular, S = skewed, N = normal, L = leptokurtic, ES = extreme skew, E-S = extreme negative skew, B = bimodal, M = multimodal, SP = spiked, T = triangular, π = dichotomous

Figure 12.2 from Glass & Hopkins, *Statistical Methods in Education and Psychology* , 3rd Edition



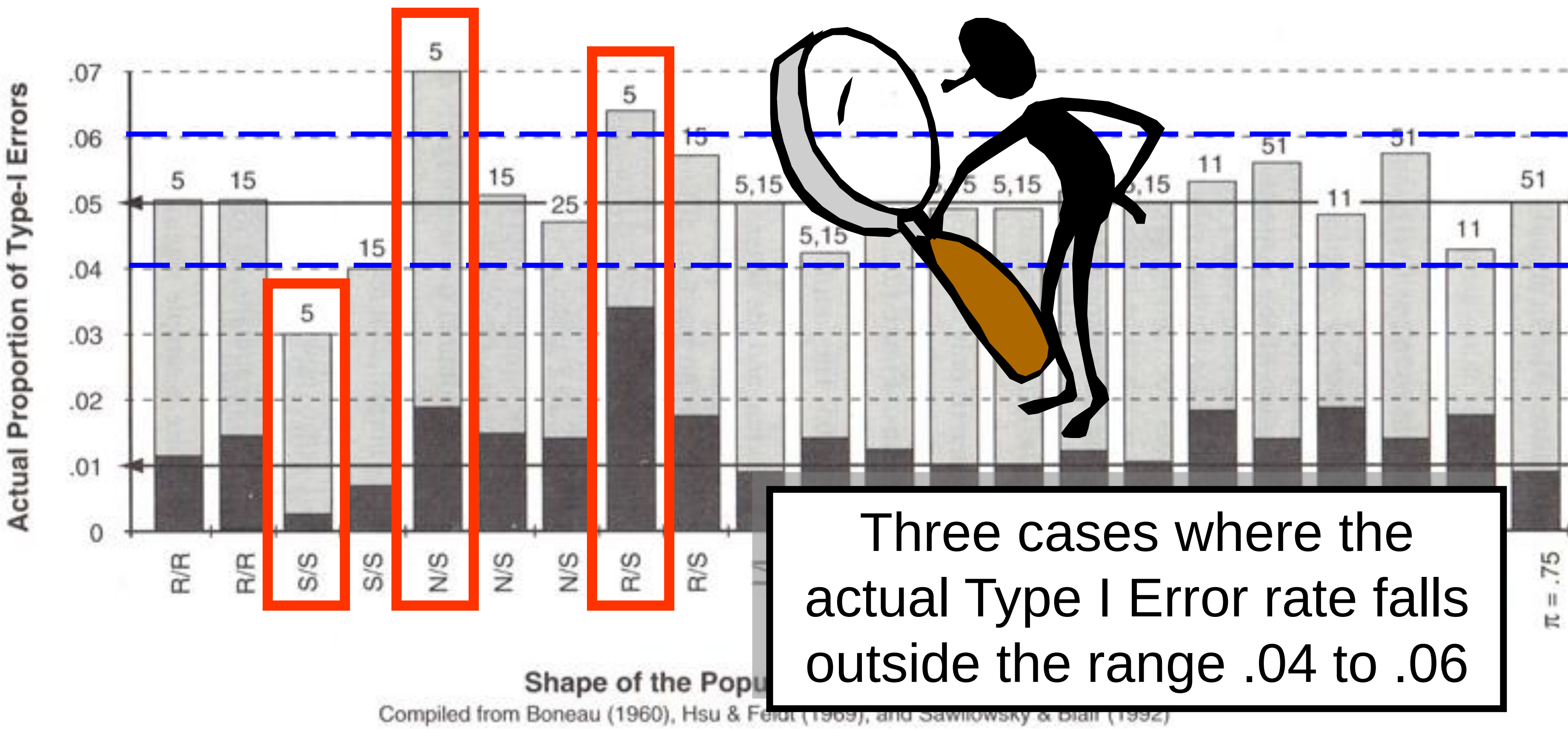
The actual proportion of Type I Errors (over repeated samples)

Figure 12.2 from Glass & Hopkins, *Statistical Methods in Education and Psychology* , 3rd Edition

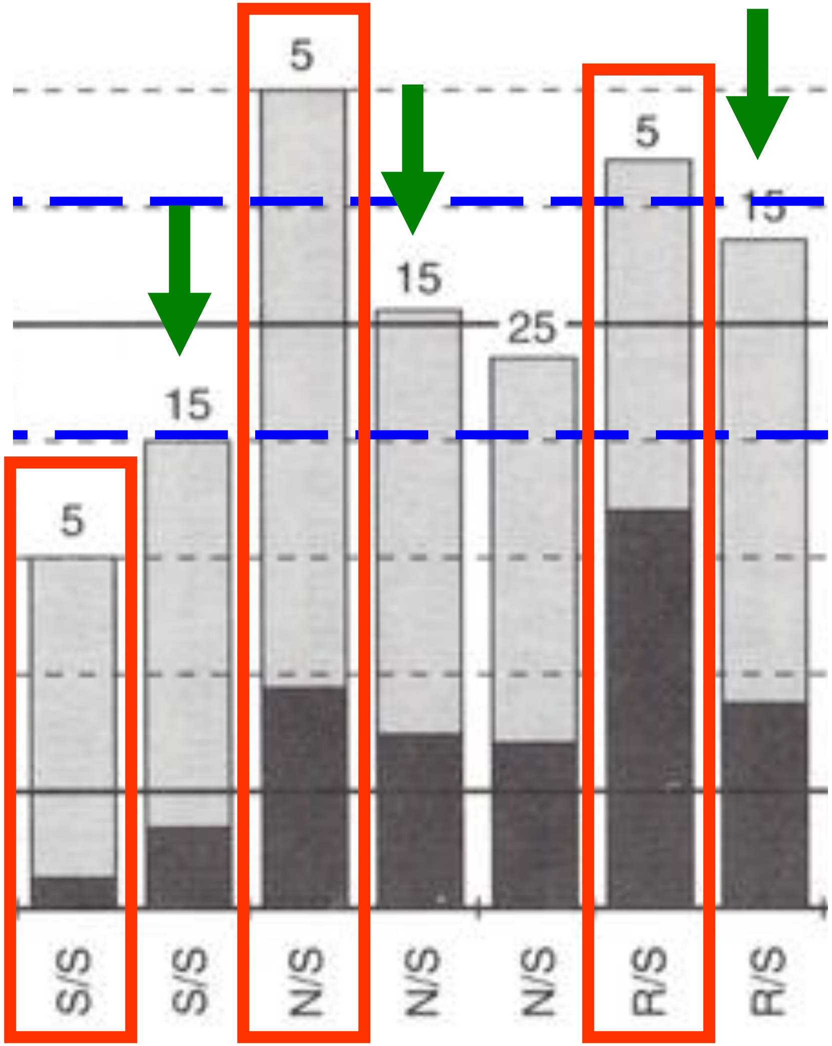


Numbers above the bars are the sample sizes; if only one number appears, both samples were the same size

Figure 12.2 from Glass & Hopkins, *Statistical Methods in Education and Psychology* , 3rd Edition



Cochran's criterion of acceptable test performance with alpha set to .05 = actual Type I error rate of .04 to .06



- S/S – both populations skewed
- N/S – one population is normal, the other skewed
- R/S – one population is rectangular, one is skewed
- In all 3 cases, $n_1 = n_2 = 5$
- In all 3 cases increasing the sample size to 15 (one bar to the right) results in test performance that meets Cochran's criterion

A List of Cases Where Cochran's Criterion was Met

- **R/R** – both populations rectangular; $n_1 = n_2 = 5$
- **S/S** – both populations skewed; $n_1 = n_2 = 15$
- **N/S** – one population normal, one skewed; $n_1 = n_2 = 15$
- **R/S** – one population rectangular, one skewed; $n_1 = n_2 = 15$
- **L/L** – both populations leptokurtic (i.e., tails thicker than the normal distribution); $n_1 = 5$, $n_2 = 15$
- **ES/ES** – both populations extremely skewed in same direction; $n_1 = 5$, $n_2 = 15$
- **M/M** – both populations **multimodal**; $n_1 = 5$, $n_2 = 15$
- **SP/SP** – both populations spiked; $n_1 = 5$, $n_2 = 15$
- **T/T** – both populations triangular; $n_1 = 5$, $n_2 = 15$

Sampling from ***Dichotomous*** Populations

- Cochran's criterion (i.e., Type I error rate between .04 and .06) was also met when samples were drawn from dichotomous populations with the following properties:
  - P = .5, Q = .5, n = 11
  - P = .6, Q = .4, n = 11
  - P = .75, Q = .25, n = 11

P and Q represent the proportions falling in the two categories

- If P and Q get too extreme (e.g., outside the range .2 to .8), test performance deteriorates

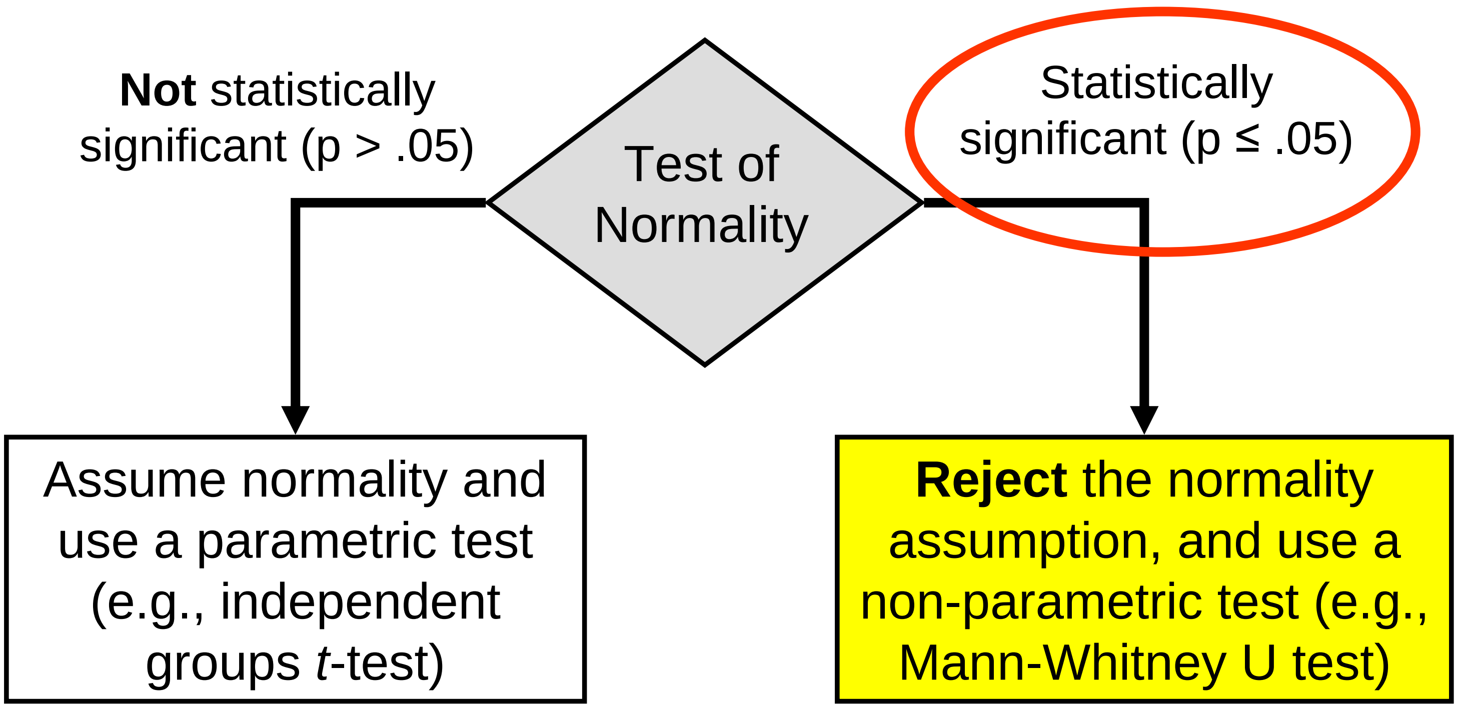
Summary of Main Points

Summary of Main Points (1)

- Textbooks list the following assumptions for t-tests:
  - Sampling from normal populations
  - Homogeneity of variance
  - Independence of observations
- The normality assumption is often listed first
- This leads (some) people to conclude that it is the most important assumption

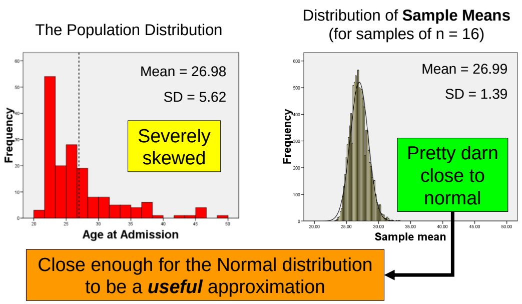
Summary of Main Points (2)

- Some textbook authors recommend using a test of normality prior to running a t-test – e.g., Field (2000) recommended the procedure shown below:



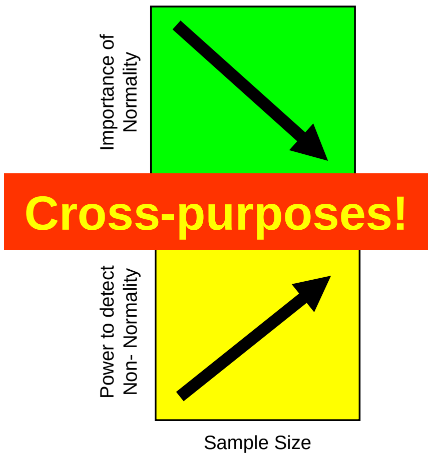
Summary of Main Points (3)

- But the important thing is normality of the **sampling distribution** of the statistic in the numerator of the t-ratio



- As the sample size increases, that sampling distribution converges on the normal distribution, ***regardless*** of population shape

- But at the same time, tests of normality become more and more ***powerful*** – i.e., they are more and more likely to detect departures from normality as those departures become less and less important for the validity of the t-test



Therefore, testing for normality prior to running a t-test is rather **silly** and **pointless**.

I agree!

And so…

do we.

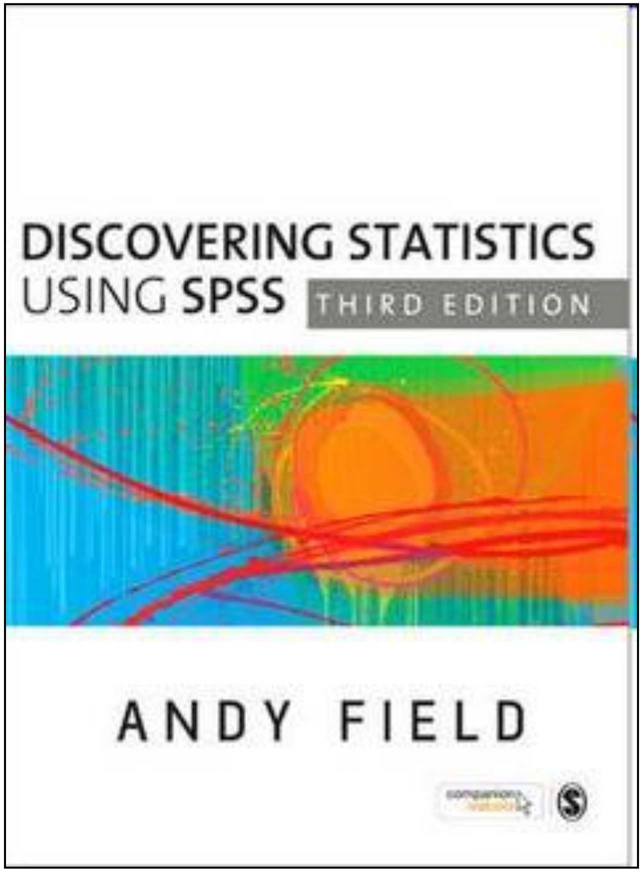
If not testing for normality, then what?

Yours truly

If means and standard deviations are sensible and appropriate for **description**, then t-tests (or ANOVA etc) will likely be just fine for **inference**.

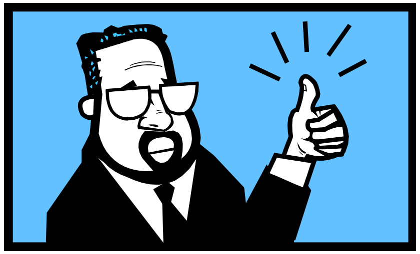
E.g., reasonably symmetrical distribution with no outliers or extreme scores

Apologies to Andy Field



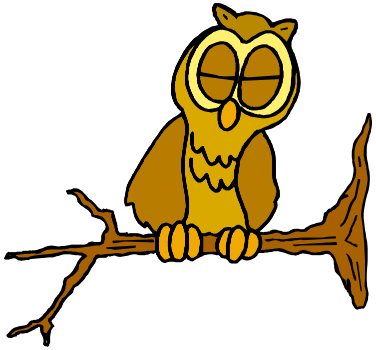
Third Edition (2009)

- Finally, in case Andy Field's lawyer is present, let me point out that the bad advice about testing for normality given in Field (2000) does **not** appear in the third edition of the book (Field, 2009).



A Final Disclaimer

No animals, *cartoon or real*, were harmed during the production of this presentation.



Okay…it's over!

Time to wake up!

Any Questions?

bweaver@lakeheadu.ca

Go see our posters!

Bruce Weaver

E-mail: bweaver@lakeheadu.ca

Tel: 807-346-7704

CENTRE FOR RESEARCH ON CRSD SAFE DRIVING

Lakehead UNIVERSITY

LAKEHEAD · LAURENTIAN · MEDICINE
Northern Ontario School of Medicine

References

Cochran WG. The χ^2 correction for continuity. *Iowa State College Journal of Science* 1942; **16**:421–436

Dawson B, Trapp RG. (2004). *Basic and clinical Biostatistics* (4th Ed.). New York, NY: Lange Medical Books / McGraw-Hill.

Field A. (2000). *Discovering Statistics using SPSS for Windows*. London: Sage.

Field A. (2009). *Discovering Statistics using SPSS for Windows (and sex and drugs and rock 'n' roll)* (3rd Ed). Los Angeles, CA: Sage.

Glass GV, Hopkins KD. (1996). *Statistical Methods in Education and Psychology* (3rd Ed). Boston, MA: Allyn and Bacon.

The Cutting Room Floor

Screening for Medically At-Risk Drivers with the SIMARD-MD: A Failure to Apply the CIHR Knowledge-to-Action Framework

Michel Bédard, PhD[1-3]; David B. Hogan, M.D[4].; Bruce Weaver, MSc[1,2]
1. Lakehead University; 2. Northern Ontario School of Medicine; 3. St. Joseph's Care Group; 4. University of Calgary

Summary of Key Points

1. British Columbia has adopted the SIMARD-MD for screening individuals who "may have cognitive impairment or dementia that could affect their ability to drive".

2. The first step in the CIHR Knowledge-to-Action framework is ***synthesis***—i.e., systematic review and critical appraisal of the existing evidence. This requires the existence of *multiple independent studies*.

3. Currently, only one published study examines the use of the SIMARD-MD as it is being implemented in BC. Therefore, synthesis is not possible, and the CIHR Knowledge-to-Action Framework **has not been followed**.

Background

2010 BC Guide in Determining Fitness to Drive

Suspected cognitive impairment that may affect ability to drive → Screen with SIMARD-MD → Score ≤ 70 → DriveABLE

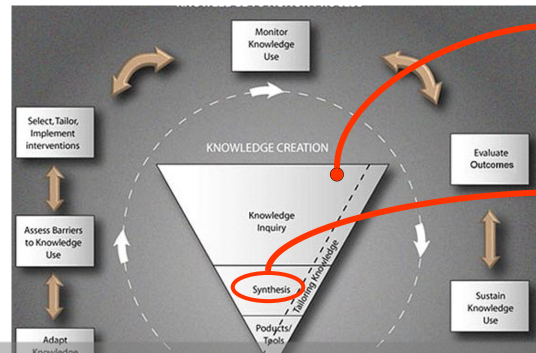
(**S**creening Tool for the **I**dentification of **M**edically **A**t-**R**isk **D**rivers: **M**odification of the **D**emTect)

(A **for-profit**, University of Alberta "spin-off" company that provides both cognitive testing and on-road driving assessments)

- The *2010 BC Guide in Determining Fitness to Drive* is published by the BC Office of the Superintendent of Motor Vehicles (OSMV)
- It can be downloaded from the BCMA website (https://www.bcma.org/publications-media/handbooks-guides)

Failure to Apply the Knowledge-to-Action Framework

The CIHR Knowledge-to-Action Process



Currently, there is only one published study examining the SIMARD-MD.

The Introduction of a New Screening Tool for the Identification of Cognitively Impaired Medically At-Risk Drivers: The SIMARD A Modification of the DemTect

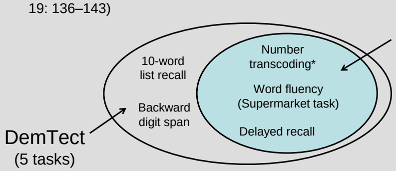
Bonnie M. Dobbs, PhD[1] and Donald Schopflocher, PhD[2]

There is nothing to synthesize. So the BCMA and the OSMV have stumbled badly at the first hurdle in the Knowledge-to-Action process.

How strong is the evidence in that one study?

GO SEE OUR POSTER!

impairment and early dementia (Kalbe et al., *Int J Geriatr Psychiatry* 2004; 19: 136–143)



Relationship Between the SIMARD-MD and DriveABLE™

The following is an excerpt from Dobbs & Schopflocher (2010, p. 126).

Declaration of Conflicting Interests

The outcome measure used in the original and validation studies (DriveABLE™ On-Road Evaluation) is part of the DriveABLE™ Assessment. DriveABLE™ Assessment Centres is a University of Alberta spin-off company. The CEO and President of DriveABLE™ Assessment Centres, Dr Allen Dobbs, is the spouse of the first author (B.D.). B.D. has no shares in or financial relationship to DriveABLE™ Assessment Centres. Dr Allen Dobbs was not involved in this research. D.S. declares no competing interests.

Definition of Knowledge Translation

"At CIHR, knowledge translation (KT) is defined as a dynamic and iterative process that includes ***synthesis***, dissemination, exchange and ethically-sound application of knowledge to improve the health of Canadians, provide more effective health services and products and strengthen the health care system." (emphasis added)

Source: http://www.cihr-irsc.gc.ca/e/39033.html

Lakehead University · CRSD Centre for Research on Safe Driving · Northern Ontario School of Medicine · St. Joseph's Care Group · University of Calgary · Canada Research Chairs / Chaires de recherche du Canada

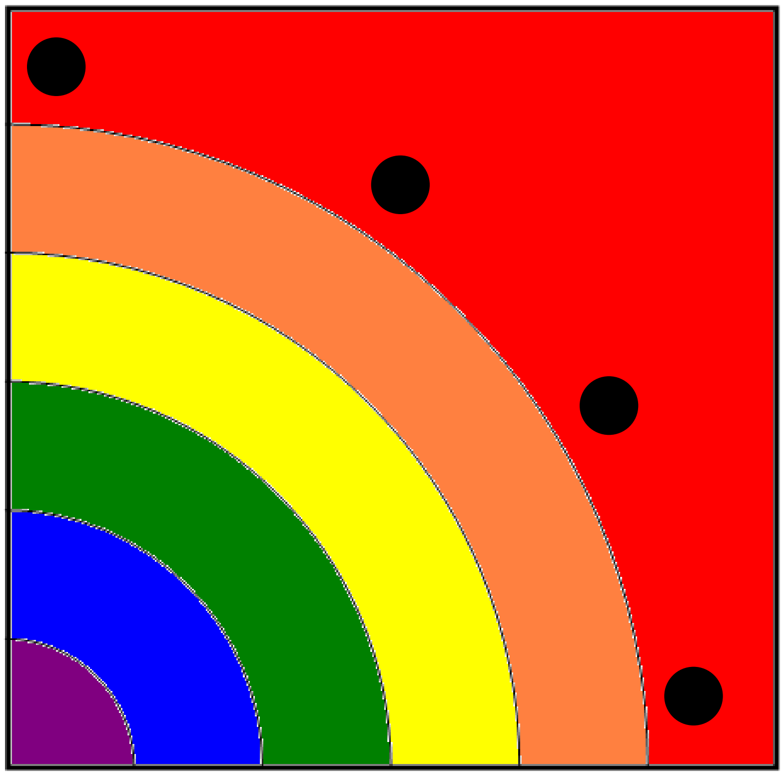
mbedard@lakeheadu.ca

The Shape of the Sampling Distribution is a Function of Population Shape and Sample Size

<u>Sample Size</u>

Large

Small



Non-Normal → Normal

Population Distribution

The colour in the plot area represents the shape of the sampling distribution of the statistic

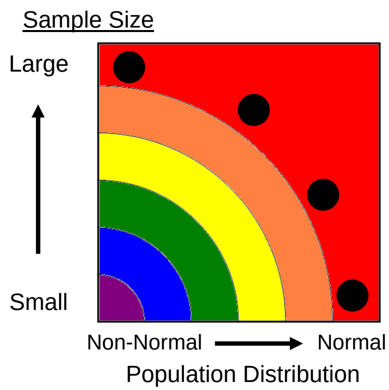
Non-Normal

Normal

Many combinations of population shape and sample size result in the same sampling distribution shape

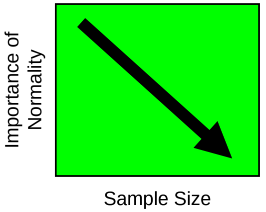
Summary of Main Points (3)

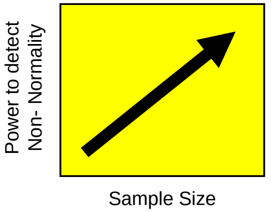
- But the important thing is normality of the **sampling distribution** of the statistic in the numerator of the t-ratio



- As the sample size increases, that sampling distribution converges on the normal distribution, ***regardless*** of population shape

- But at the same time, tests of normality become more and more ***powerful*** – i.e., they are more and more likely to detect departures from normality as those departures become less and less important for the validity of the t-test

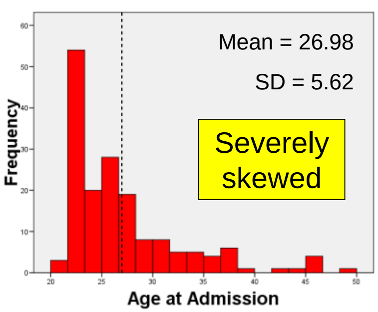




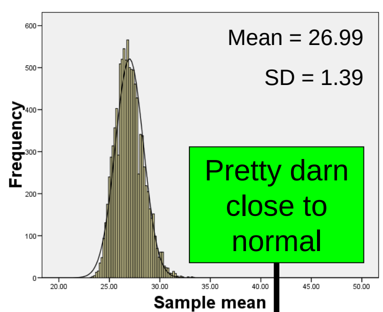
Does *n* have to be ≥ 30?

- Some books say that we should (or even must) have n ≥ 30 to ensure that the sampling distribution of the mean is approximately normal
- But the examples shown earlier demonstrate that the sampling distribution of the mean often becomes nice and symmetrical with sample sizes much lower than 30

The Population Distribution

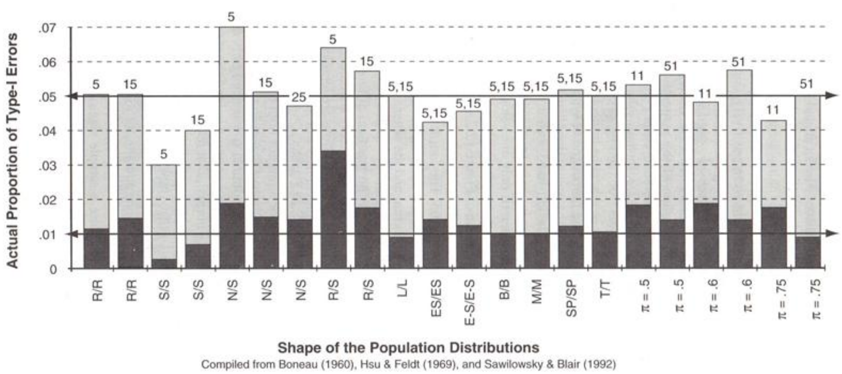


Distribution of **Sample Means**
(for samples of n = 16)



Close enough for the Normal distribution to be a ***useful*** approximation

Figure 12.2 from Glass & Hopkins,
Statistical Methods in Education and Psychology , 3rd Edition



What is the “rule of 30” about then?

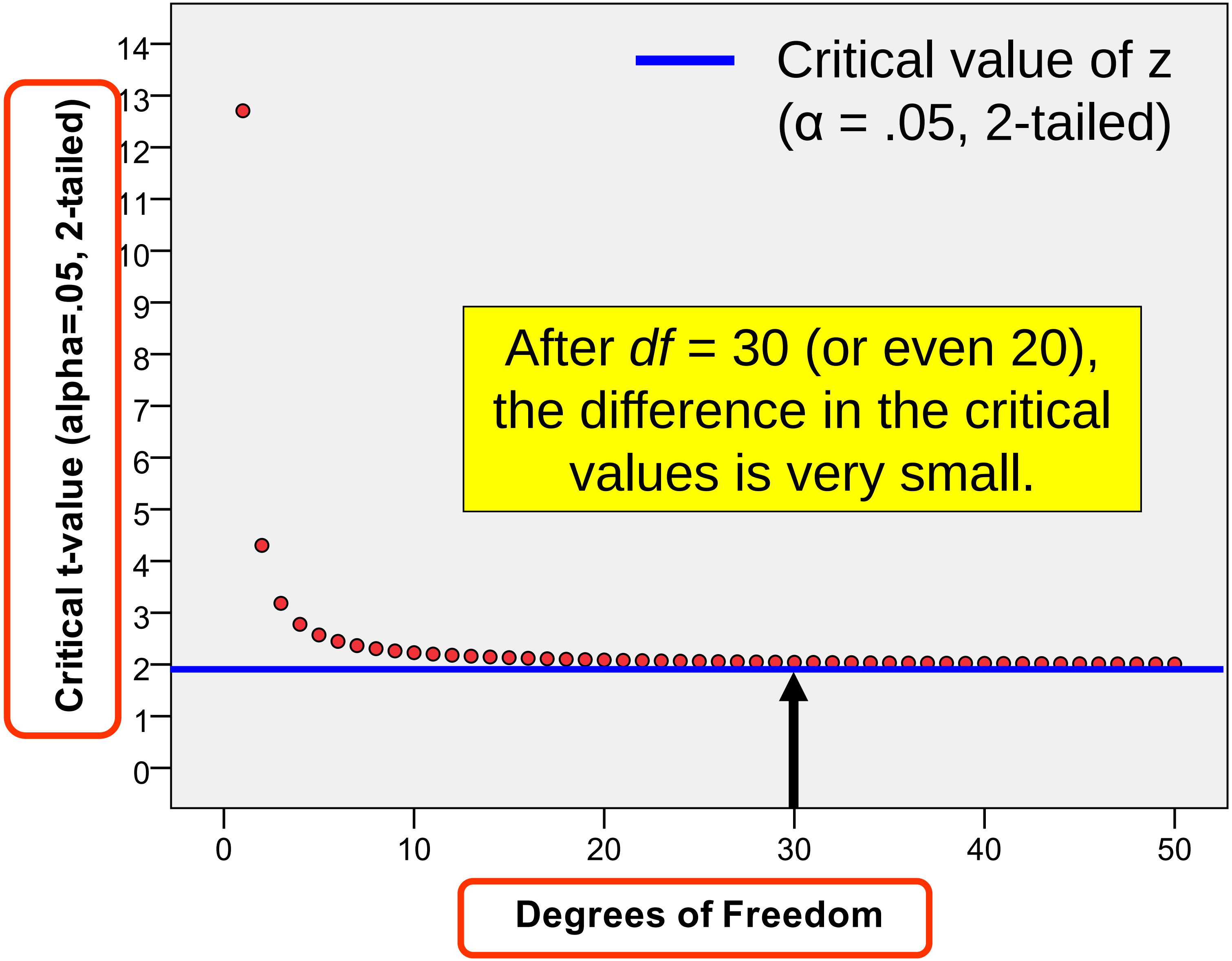
- In the olden days, textbooks often described inference for *small samples* and inference for *large samples*
- E.g., comparing the means of 2 independent samples:
  - **Small samples**: independent groups t-test using **critical value of *t***
  - **Large samples**: independent groups t-test using **critical value of *z***

Often defined as n ≥ 30

From the Standard Normal Distribution

Why make the distinction?

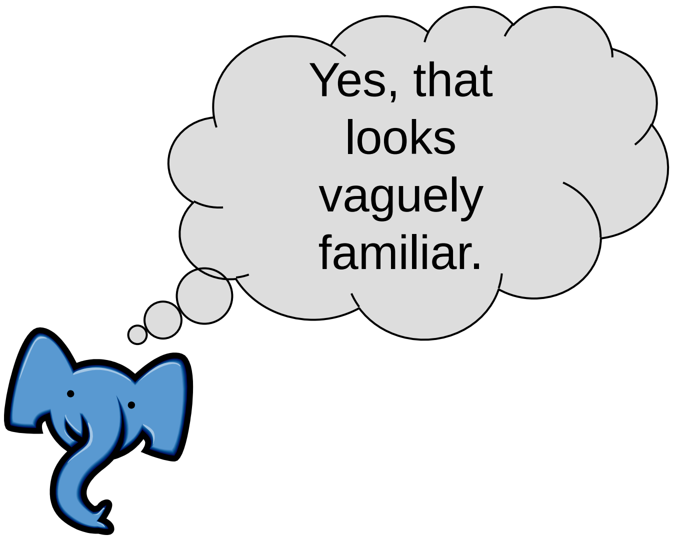
- Why did textbook authors make the distinction between small and large samples?
- Remember that in those days, data analysts used tables of critical values to determine if a test result was statistically significant
- **Tables of critical values take up a lot of room!**
- When $n \geq 30$, the critical value of z (from the Standard Normal distribution) was judged to be close enough to the critical value of t that it could be used instead
- In older books, tables of critical t-values only go up to df=30 or so


Critical value of z
(α = .05, 2-tailed)
After *df* = 30 (or even 20), the difference in the critical values is very small.
Critical t-value (alpha=.05, 2-tailed)
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0
10
20
30
40
50
Degrees of Freedom

Back to Stats Class!

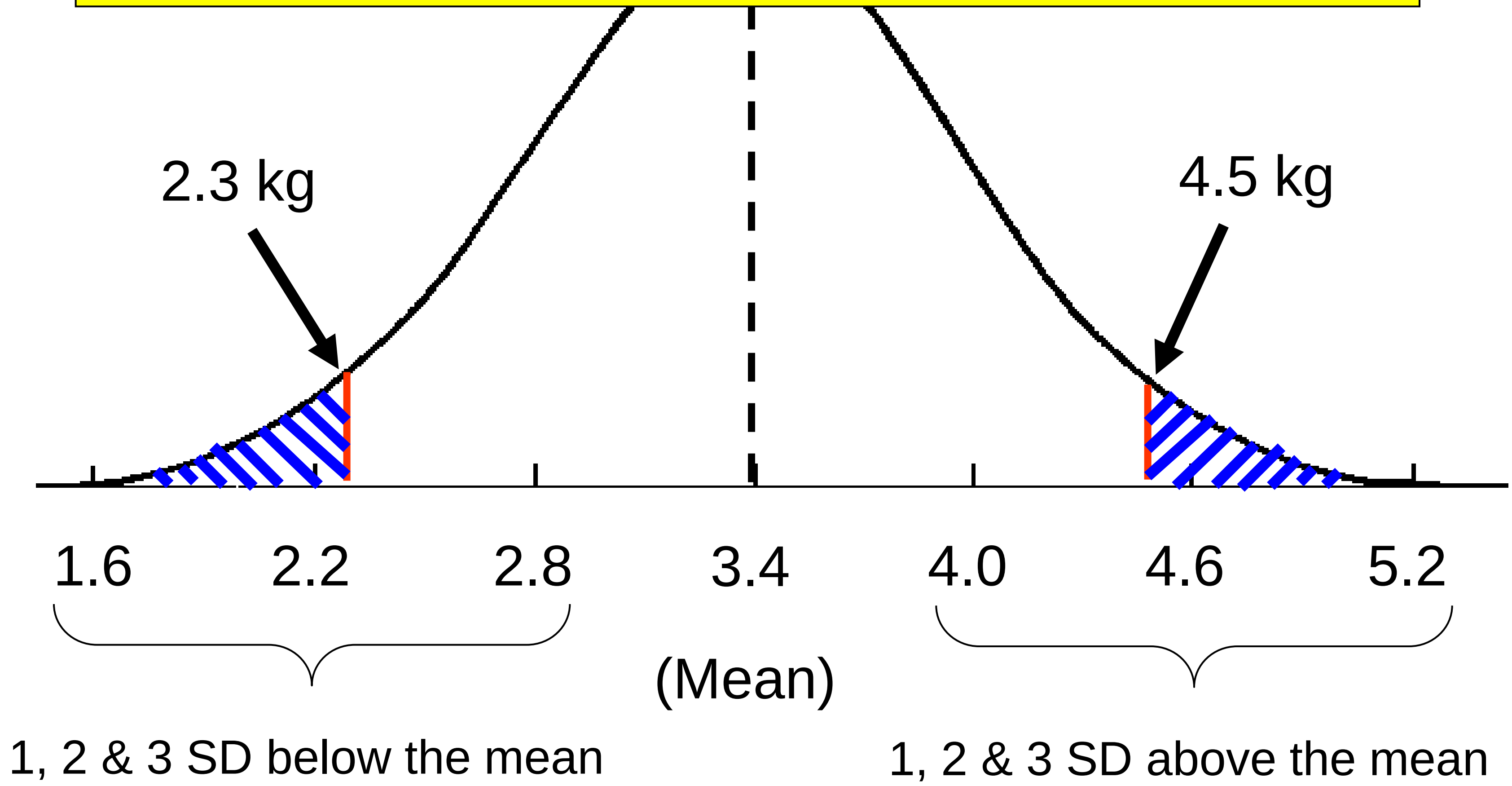
- Anyone who has taken an introductory stats class no doubt remembers tackling problems like this:

> The birth-weight of newborns in a particular hospital is (approximately) ***normally distributed*** with a ***mean of 3.4 kg*** and a ***standard deviation of 0.6 kg***. **What proportion of newborns in this population have a birth-weight ≥ 4.5 kg or ≤ 2.3 kg**?



Step 1: Sketch the Distribution

ANSWER: Area below 2.3 kg + area above 4.5 kg



Step 2: Convert to Z-scores

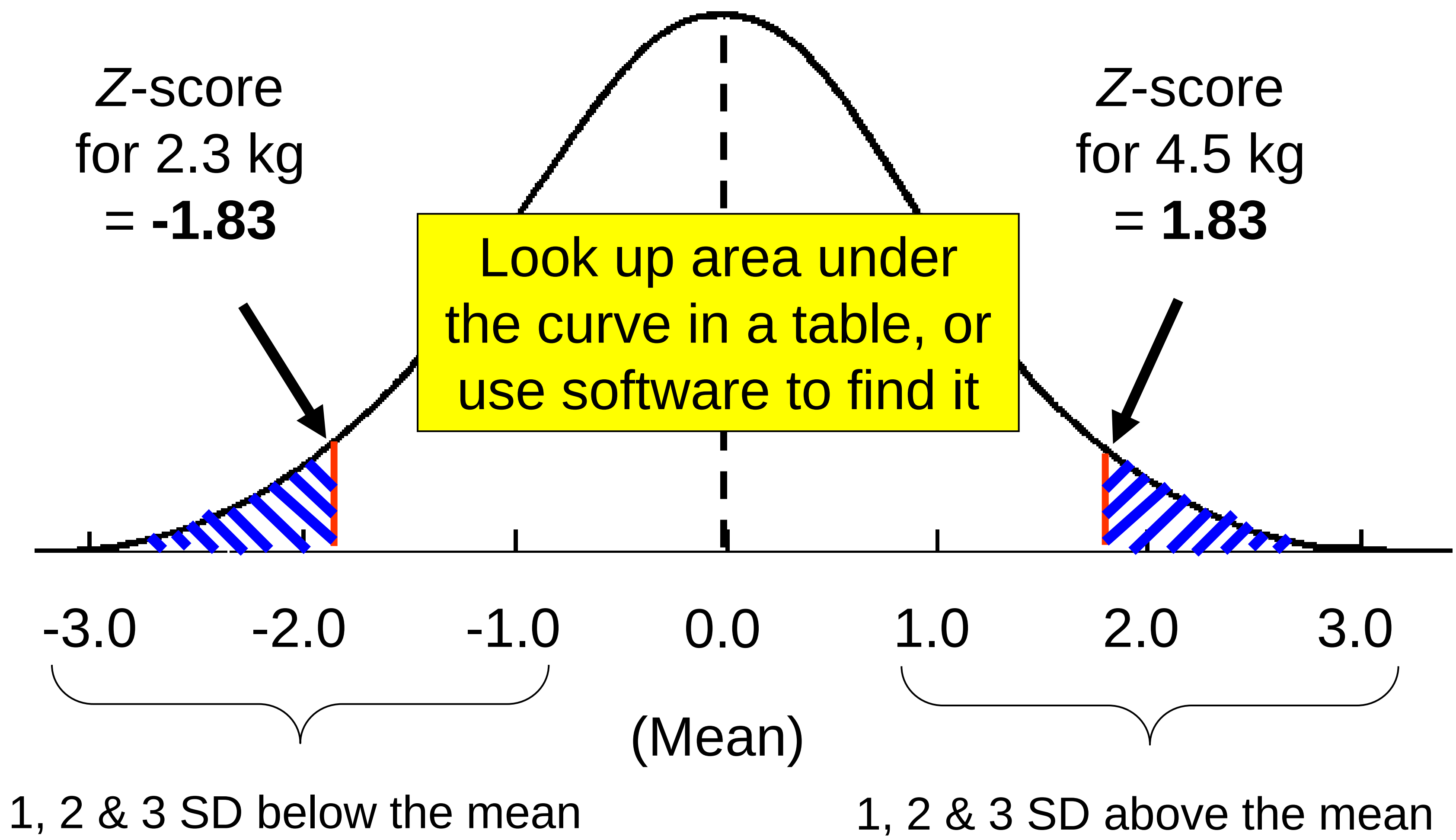
$$Z = \frac{\text{Score} - \text{Mean}}{\text{SD}}$$

Now refer these values to the standard normal distribution.

$$Z_1 = \frac{2.3 - 3.4}{0.6} = \boxed{-1.83}$$

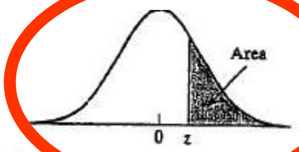
$$Z_2 = \frac{4.5 - 3.4}{0.6} = \boxed{1.83}$$

Step 3: Sketch the Standard Normal Distribution



Step 4 the old fashioned way

Table 4. Normal curve areas
Standard normal probability in right-hand tail
(for negative values of z areas are found by symmetry)



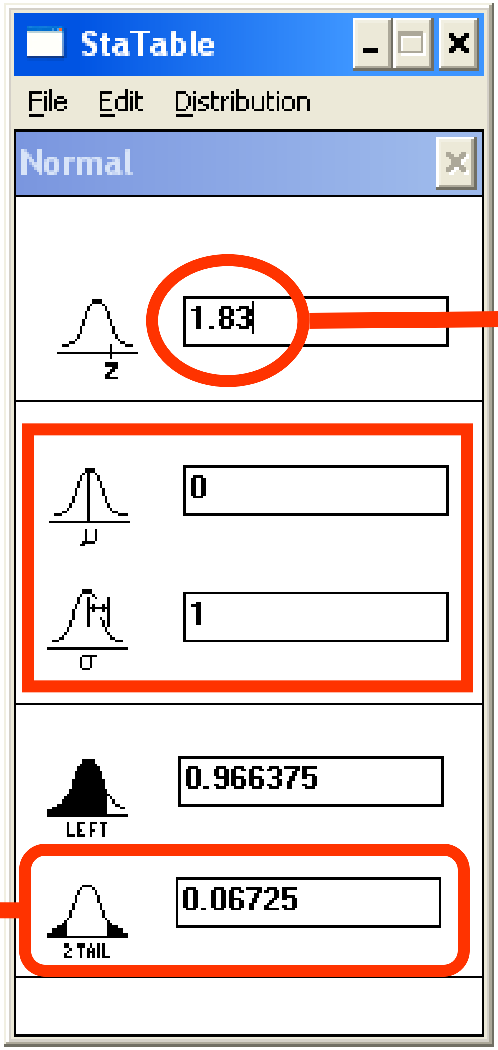
	Second decimal place of z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2843	.2810	.2776
0.6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2514	.2483	.2451
0.7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2206	.2177	.2148
0.8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1922	.1894	.1867
0.9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1660	.1635	.1611
1.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1423	.1401	.1379
1.1	.1357	.1335	.1314	[illegible]	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	[illegible]	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	[illegible]	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	[illegible]	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	[illegible]	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	[illegible]	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	[illegible]	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	[illegible]	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110

Entry for *Z*-score of 1.83 =**.0336**

- Area above 1.83 = .0336
- Normal distribution is symmetrical about the mean
- Therefore, the area **below** -1.83 = .0336 too
- Therefore, the proportion of newborns having a birth-weight ≥ 4.5 kg or ≤ 2.3 kg is .0336 2 = **.0672**, or **6.72%**

Step 4 the new-fangled way

- Nowadays, one would probably use **software** to obtain the area ≤ *Z* = -1.833 plus the area ≥ *Z* = 1.833
- E.g., StaTable from www.cytel.com



Z = 1.83

Standard normal (μ=0, σ=1)

Area ≤ -1.83 plus area ≥ 1.83 = **.06725**

Why Did That Work?

- Z-score problems like that worked because the distribution of scores was (approximately) normal
- In that case, you can transform to Z-scores and refer to the Standard Normal distribution

The Normal Distribution

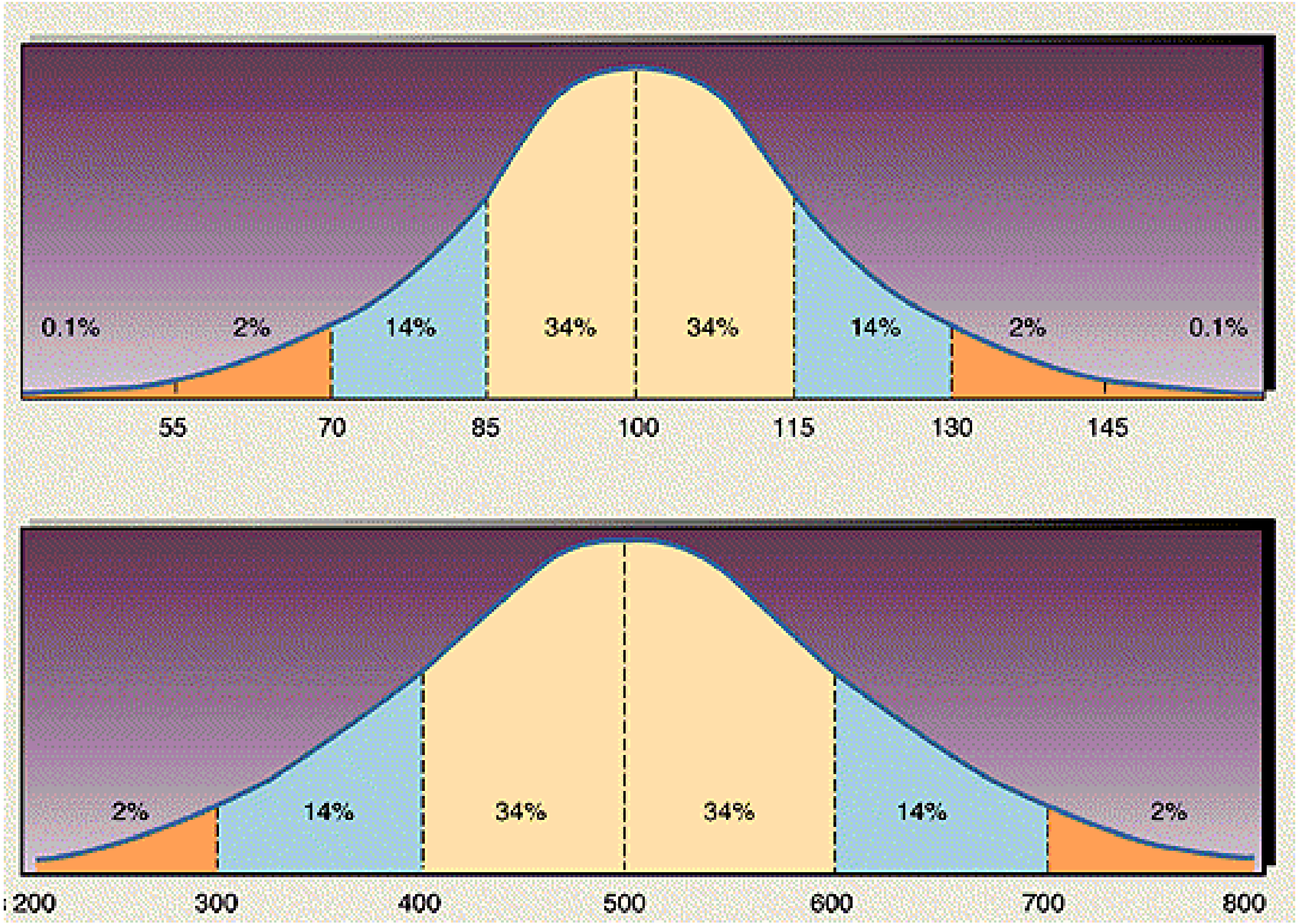
- Anyone who has taken an introductory stats course will remember the ***Normal*** (bell-shaped) distribution
- Actually, a ***family*** of Normal distributions

Mean

$$f(y) = \frac{e^{-(Y-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$$

Standard Deviation

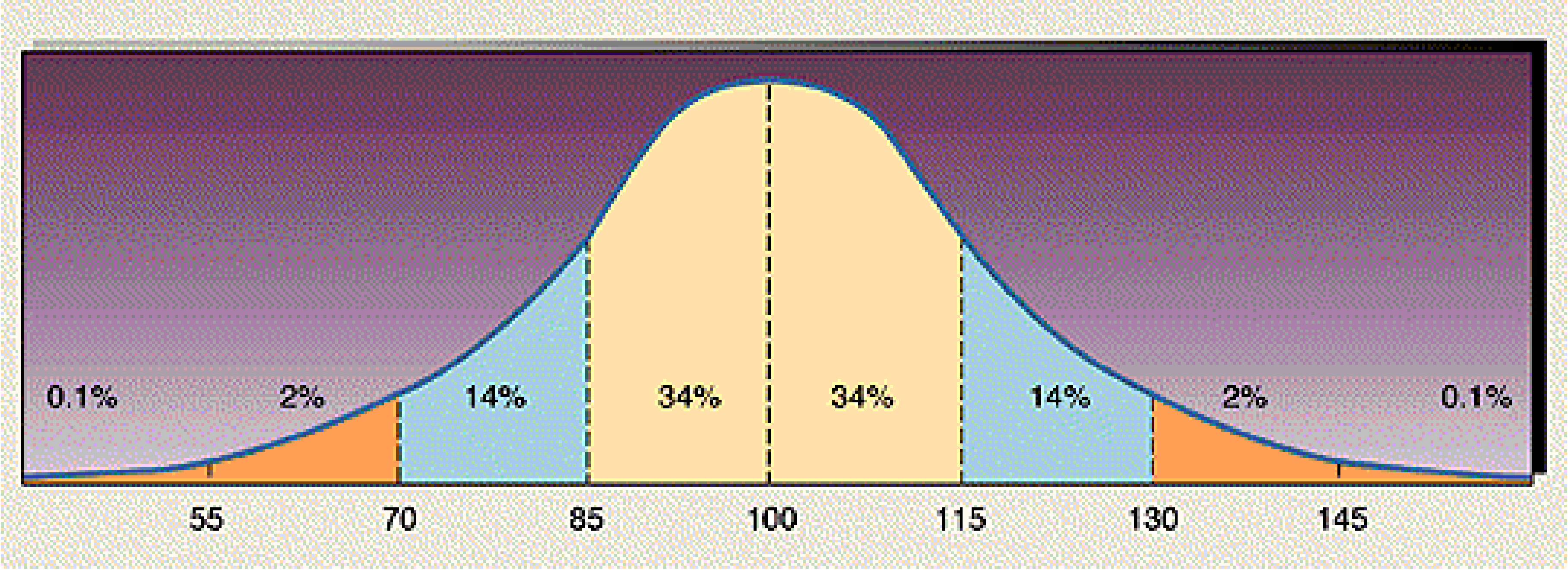
Some Examples



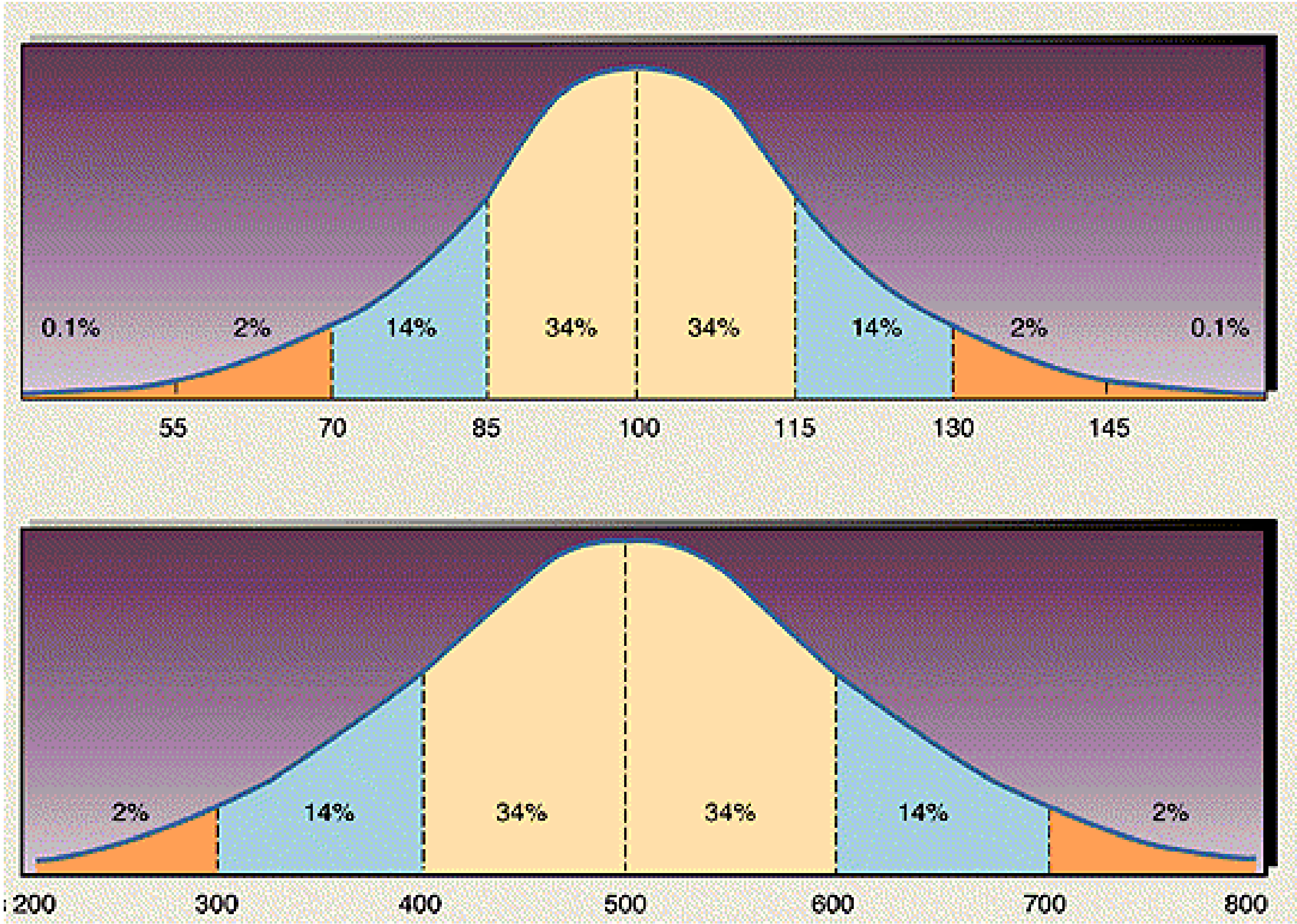
$\mu = 100$

$\sigma = 15$

Example 1: $\mu = 100, \sigma = 15$



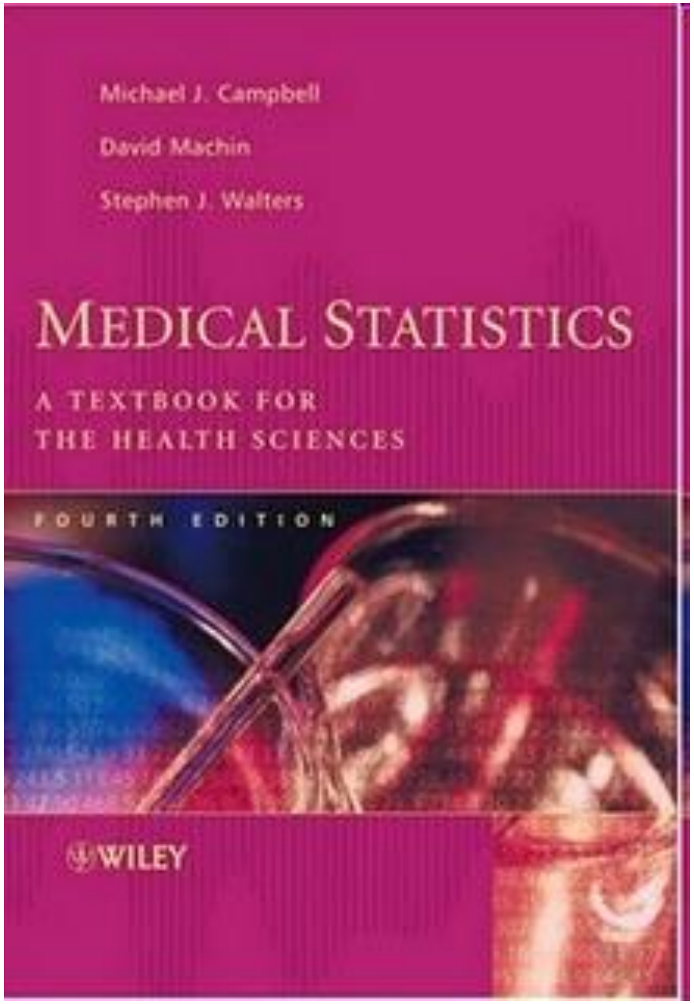
Some Examples



μ =100

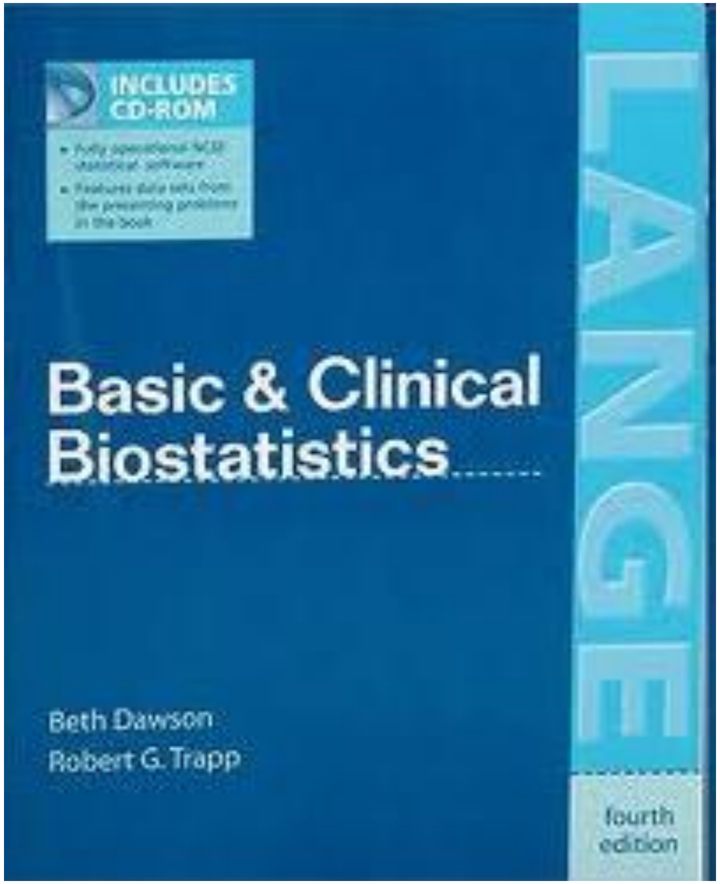
σ = 15

Assumptions for the Unpaired t-test (x)



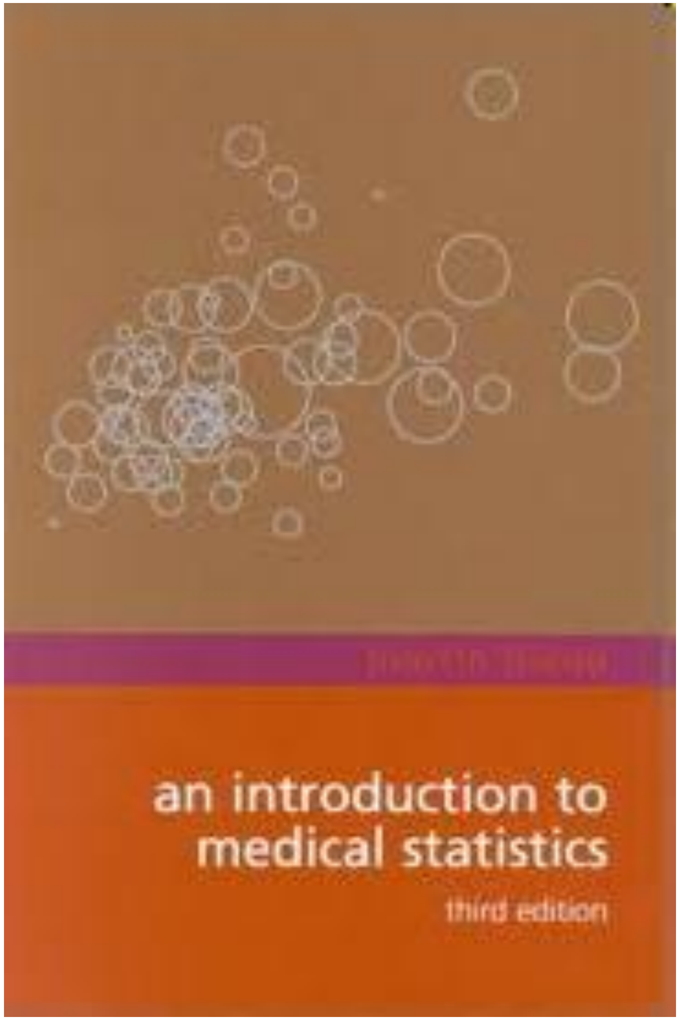
1. The groups are independent.
2. The [dependent] variables of interest are continuous.
3. The data in both groups have similar standard deviations.
4. The data is Normally distributed in both groups.

Assumptions for the Unpaired t-test (x)



1. "…the observations in *each group* follow a normal distribution."
2. "The standard deviations (or variances) in the two samples are assumed to be equal"
3. "…independence, meaning that knowing the values of the observations in one group tells us nothing about the observations in the other group"

Assumptions for the Unpaired *t*-test (x)



- "Not only must the samples be from Normal distributions, they must be from Normal distributions with the same variance."
- Also clear from the section heading that the two samples must be independent

How z- and t-tests **really** work (2)

$$z \text{ or } t = \frac{\text{statistic - parameter}|H_0}{\boxed{SE_{statistic}}}$$

- If the population SE is known, this ratio = Z, and the standard normal distribution can be used
- If the population SE is **not** known, it must be estimated (using the sample standard deviation)
- In that case, ratio = *t*, and its sampling distribution is a *t*-distribution with appropriate degrees of freedom (df)

How z- and t-tests **really** work (3)

<u>Single-sample z-test</u>

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

Statistic

Parameter under True H_0

SE of the statistic in the numerator

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}$$

Exact vs. Approximate Tests

- A test is <u>exact</u> if the sampling distribution of the test statistic is given **exactly** by the mathematical distribution used to obtain the p-value
- E.g., the binomial distribution gives exactly the sampling distribution of X (the number of Heads) in coin-flipping experiments
- A test is <u>approximate</u> if the mathematical distribution only **approximates** the true sampling distribution of the test statistic
- E.g., chi-square tests are approximate tests

Back to the Assumptions

- As we've seen, textbooks often list the assumptions for a *t*-test as something like this:

1. The data must be sampled from a **normally distributed population** (or populations in case of a two-sample test).

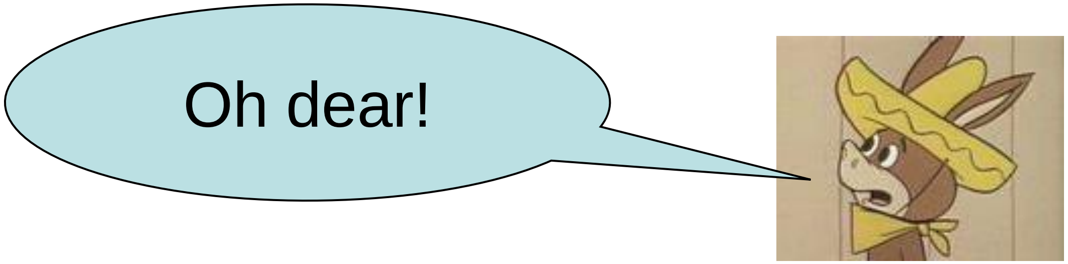
2. For two-sample tests, the two populations must have **equal variances**.

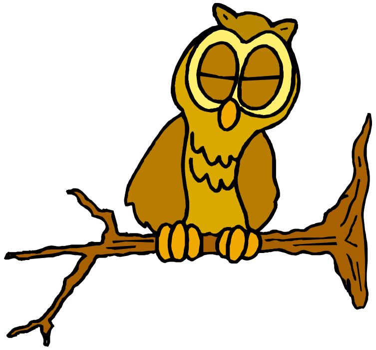
3. Each score (or difference score for the paired t-test) must be **independent** of all other scores.

The First Two Assumptions are Never Met

1. As Box noted, nothing in nature is truly normal
2. Furthermore, it is virtually impossible for two different populations to have variances that are identical down to the last decimal place.

Therefore, no one who is working with real data meets the assumptions of normality and homogeneity of variance.





Okay…it's over!

Time to wake up!

Any Questions?

bweaver@lakeheadu.ca