Section IV

Sampling distributions

Confidence intervals

Hypothesis testing and p values

Section IV

IV- Sampling distributions, confidence intervals and hypothesis testing

Sampling distributions and the Central Limit Theorem

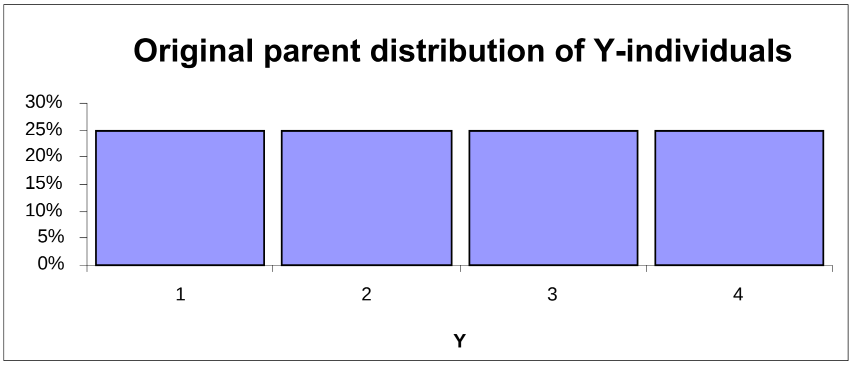
For all but very small samples, sample means and other summary statistics (medians, SDs, proportions, odds ratios, risk ratios,correlation coefficients,…) follow the Gaussian ("normal") distribution from study to study. That is, from repeated sampling from the same target population, the distribution of the sample summary statistic tends toward the bell shaped Gaussian with the mean centered on the true population value. This fact is called the **central limit theorem**.

The simplest example is the sample mean, $\bar{Y}$. If individual measurements Y have mean μ and standard deviation σ, then one can show that the "sampling" distribution of $\bar{Y}$ is Gaussian with mean μ and SD equal to σ/√n. This is true for $\bar{Y}$ no matter what distribution Y has. The quantity σ/√n is called the **standard error** (SE) of $\bar{Y}$, to distinguish it from the standard deviation (SD) of Y. (Don't get those two mixed up!). Note that the SE, (but not the SD) depends on **n**, the sample size, and the SE gets smaller as n increases.

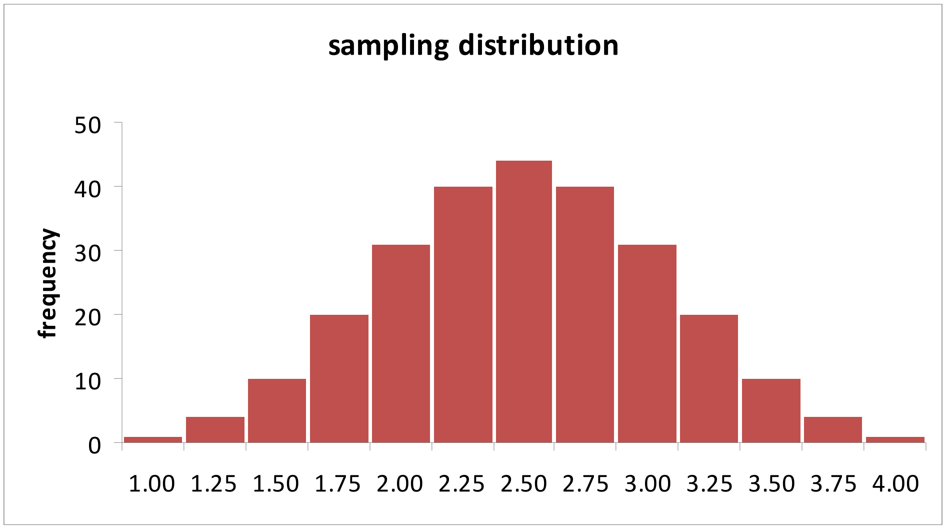
The standard error indicates how much a sample statistic varies from sample to sample around the true population value. It is not the same as an SD which indicates how much one person's value varies from person to person.

Central limit theorem demonstration - SD vs SE

Consider a population of patients recovering from disease whose recovery level, Y, is rated on a four point scale: 1=poor, 2=fair, 3=good or 4=excellent. In this example, we assume that the four points of the "Y" recovery scale are evenly spaced. In the population of all patients, assume we know that the distribution of Y is 25% 1, 25% 2, 25% 3 and 25% 4. Therefore, the population mean is μ=2.5 and the population SD is σ =1.12. The population distribution of Y is NOT Gaussian.



Now imagine taking a sample from this population of size n=4 patients. Possible samples are 3,1,2,2, ..   2,4,3,1 etc.. The sample mean of the four sample data values ranges from a mean value of 1, when the sample is 1,1,1,1 to a mean of 4 when the sample is 4,4,4,4.



Distribution of the sample means – sampling distribution

mean = 2.5, SD = 0.56, n=4

The histogram above shows the distribution of the sample means (a summary statistic) from all possible samples that could be taken from this population**. This distribution is called the sampling distribution** (of sample means, in this case). Not surprisingly, the mean of all the sample means from all possible samples (the mean of the sampling distribution) is also μ = 2.5. This does not imply that every sample has mean 2.5, only the mean of all possible sample means is 2.5. That is, **on average**, the sample mean and the population mean, μ, are the same. The SD of the sampling distribution is 0.56. This SD is not zero since not all samples have the same mean. We will call the standard deviation of the sampling distribution the **standard error (SE)**, to distinguish this from the SD of the original population values. So 0.56 is the standard error.

Interestingly, the **SE is equal to the SD of the Ys divided by √n.**

SE   =   SD/√n   the square root n law

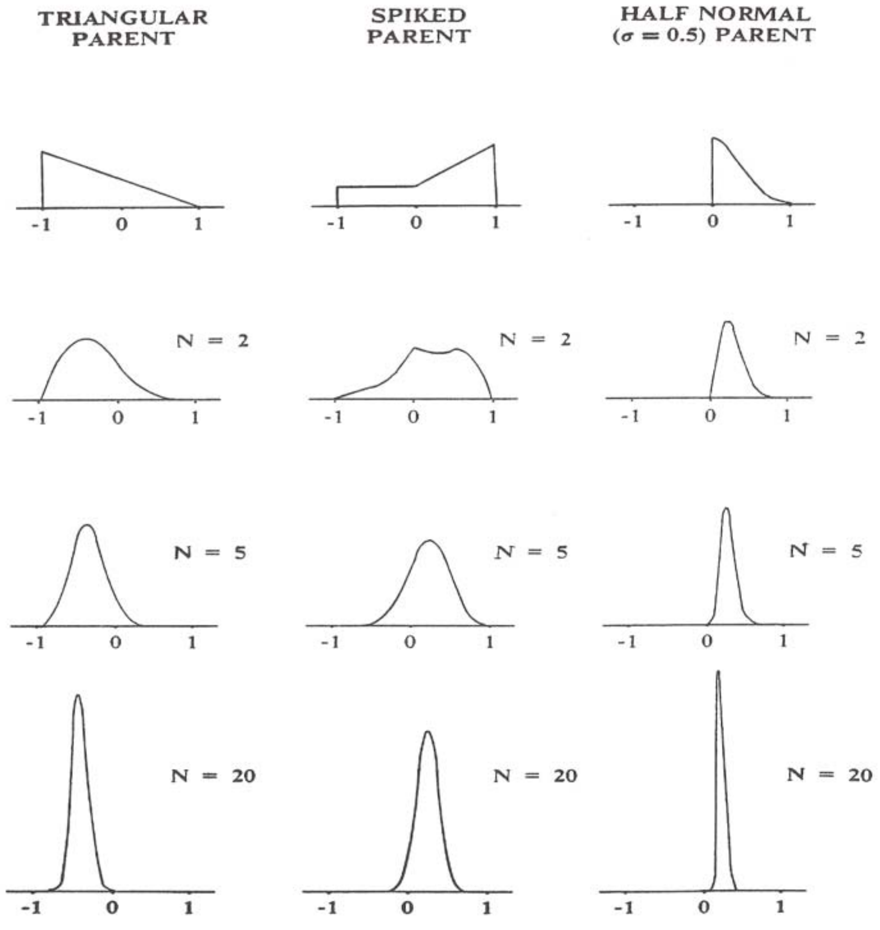
In this example, n=4, √4=2.   So 1.12 / √n = 1.12/2 = 0.56.

That is, as the sample size (n) gets larger, the "spread" of the sampling distribution (as measured by the SE) is narrower and the sample results cluster more closely around the true population value, μ. While the **SE** gets smaller as n increase, the SD does not change systematically.

The SD is a measure of the variation of Y from patient to patient. The SE, in contrast, is the measure of variation of the sample **statistic** from sample to sample. That is, the SE is a measure of how precisely the sample statistic estimates the true population parameter.

Statisticians have shown that, even if the patient measurements Y do not have a Gaussian distribution, summary sample statistics, such as sample means, have a sampling distribution that is usually well approximated by a Gaussian distribution whose mean is the same as the population parameter being estimated and whose standard deviation is equal to the standard error. This approximation gets better as n gets larger. Of course, this result is only true if all of the samples are taken at random from the same population. **This result is called the Central Limit Theorem. A rule of thumb is that the statistic (such as the mean) will almost always follow the Gaussian distribution if n is at least 30.** (If n is less than 30 and Y is not Gaussian, may need non parametric methods)

CENTRAL LIMIT THEOREM EXAMPLES ASYMMETRIC PARENT

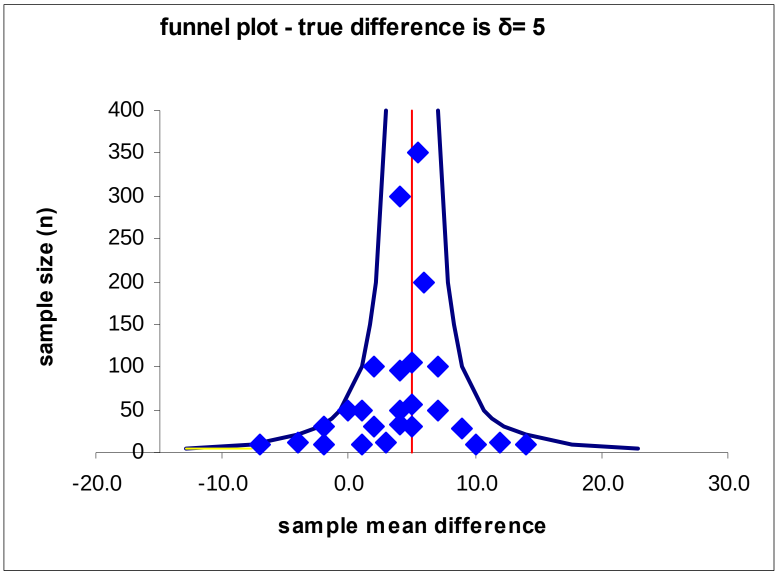


© 1993 Mosby–Year Book, Inc.

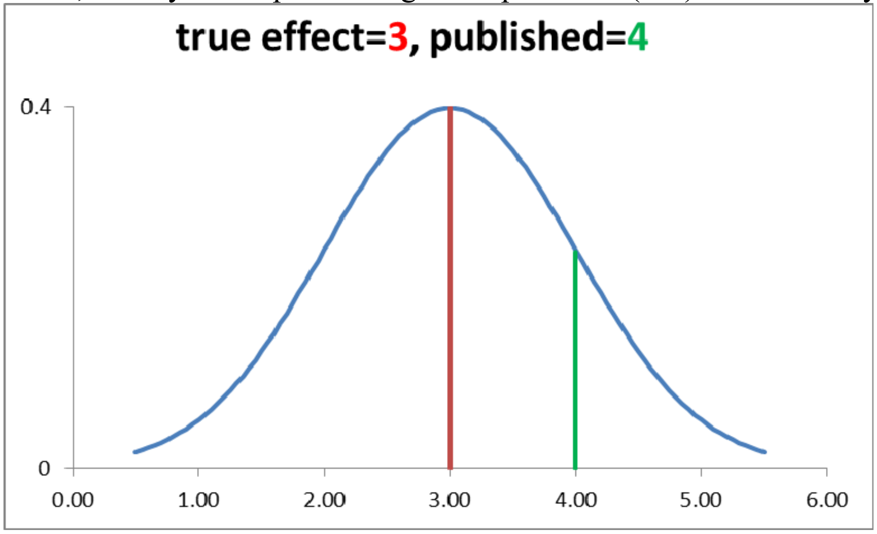
Meta – Analysis (part of a systematic review)

The Central Limit Theorem can often be observed directly when looking at the effect of the same treatment comparison replicated across many studies (ie many samples) when all the studies were carried out on the same type of population and the treatment effect is measured with the same statistic.

A funnel plot is a plot of the treatment effect (here the mean difference) versus the SE or versus n. As n, the sample size, gets larger, the statistic gets closer to the true population value.

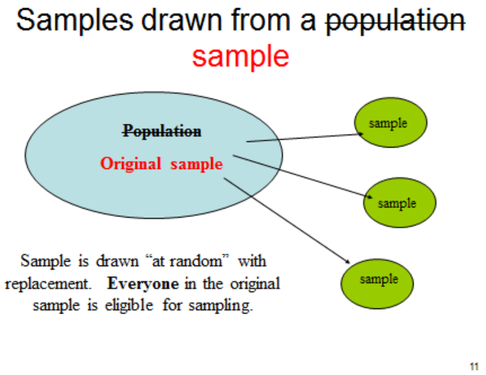


Publication bias - This also helps explain why some studies are not well reproduced. If, for example, the sample mean difference follows a normal distribution and the true value is “3”, a study that reports a larger sample effect (“4”) is more likely to be published.



Resampling – the “Bootstrap”

While one does not repeatedly sample from the same population, (that is, one only carries out the study once), a “simulation” of repeated sampling from the population can be obtained by repeatedly sampling from the **sample with replacement & computing the statistic from each resample,** creating an “estimated” sampling distribution. The SD of the statistics across all the “resamples” is an estimate of the standard error (SE) for the statistic.



Samples drawn from a ~~population~~ sample

Sample is drawn “at random” with replacement. **Everyone** in the original sample is eligible for sampling.

Confidence intervals

We can use this Gaussian distribution property of sample means, sample mean differences or any **sample statistic** to form a **confidence interval** for the true underlying population parameter in the population from which the sample is taken. That is, we can use this Gaussian property to make formal generalizations (inferences) about what the entire population is like “in general” based on only a (hopefully representative) sample. In medicine, we need to know about the overall properties of people and treatments in general (ie. in the population), not just in the one particular sample that was examined in a particular study.

Using Gaussian theory, a **confidence interval** is usually computed as

Sample Statistic +/- Z_{tabled} SE

where “sample statistic” is any summary statistic such as a mean difference or a proportion. Z_{tabled} is a specifically chosen percentile from the Gaussian table and SE is the standard error that corresponds to the summary statistic. The formula for the SE is different for different statistics and can be complicated.

The value of Z_{tabled} determines the level of confidence. Most commonly, for a 95% confidence interval, the 97.5th Gaussian percentile, Z=1.96, is chosen (so there is 2.5% in each tail). For a 90% confidence interval, the 95th Gaussian percentile is needed (5% in each tail). For an 80% confidence interval, the 90th percentile is needed etc.

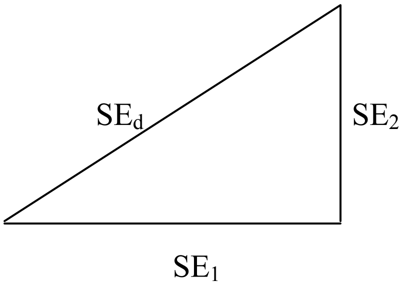
Mean differences and their SE - As a simple extension, the mean difference $\bar{d} = \bar{Y}_1 - \bar{Y}_2$ also has a Gaussian distribution. If the measurements Y_1 have mean μ_1 and SD σ_1 and Y_2 have mean μ_2 and SD σ_2, then

$\bar{Y}_1$ has mean μ_1 and SE = $\sigma_1/\sqrt{n_1} = SE_1$

$\bar{Y}_2$ has mean μ_2 and SE = $\sigma_2/\sqrt{n_2} = SE_2$

and $\bar{d}$ has mean $\delta = \mu_1 - \mu_2$ and
SE = $\sqrt{[\sigma_1^2/n_1 + \sigma_2^2/n_2]} = SE_d$

Note that $SE_d = \sqrt{SE_1^2 + SE_2^2}$



The sample statistic $\bar{d}$ has a Gaussian distribution centered at the population mean difference δ even if Y_1 and Y_2 (the individual's observations) are not Gaussian!

SE and confidence interval for a proportion

For a proportion P, the SE is sqrt[P(1-P)/n]. The approximate confidence interval is given by P +/- Z_{tabled} SE where Z is the Gaussian percentile (ie Z=1.96 for 95% confidence).

Example: IOP – HBA1c change in diabetics with baseline HBA1c > 7.5%
(Pratley et. al. Lancet, 2010, 375, 1447-56)

Statistics for HBA1c change from baseline to 26 weeks after treatment

Treatment	n	Mean change	SD	SEM
Liraglutide	225	-1.24	0.99	0.066
Sitaglipin	219	-0.90	0.98	0.066

Mean difference = $\bar{d}$ = 0.34%, $SE_d = \sqrt{0.066^2 + 0.066^2}$ = 0.093%

For 95% confidence, use 97.5th percentile
Z = 1.96 (if use t dist, df= 225+219-2=442 , t=1.97 instead of Z=1.96)

95% Confidence interval for true mean difference in mean changes is

0.34% +/- 1.96 (0.093%) or 0.34 +/- 0.182 or (0.16% , 0.52%)

We observed a mean difference of 0.34% favoring Liraglutide in this study. The confidence interval implies that the true population mean difference is at least 0.16% favoring Liraglutide and might be as large as 0.52% favoring Liraglutide . Note that a mean difference of zero (no mean difference) is **<u>not</u>** included in this confidence interval.

Technical comment: t vs Z distribution – When computing the standard error (SE) for means or mean differences, we only have the sample standard deviation (S), not the population standard deviation (σ). Technically, we should only be using Gaussian Z percentiles when the SE is computed from σ. Fortunately, since we must use S in practice (σ is unknown), a statistician named Gosset determined how to "correct" the Gaussian percentiles in order to account for using S in place of σ. The corrected percentiles are called "t" percentiles instead of Z percentiles and the tables for them are tables of the "t" distribution. The "t" percentiles depend on the sample size (via the "degrees of freedom = sample size-num parameters") and are a somewhat larger than the corresponding Z percentile. So confidence intervals based on t are somewhat larger than those based on Z, accounting for the use of S instead of the population σ. In this example, t=2.08 instead of Z=1.96 so the 95% confidence interval becomes 1.5 +/- 2.08 (0.70) or (0.044 mmHg, 2.96 mmHg).

Hypothesis testing

A very common approach used in the biomedical literature when making formal comparisons is so-called "hypothesis testing", a possibly misleading term. In the case of comparing outcomes between two groups (i.e comparing the outcomes of two medical treatments) the idea behind hypothesis testing is to make a "null" hypothesis assumption that, in truth, there is **no** (average) difference between the outcomes in the two groups in general (i.e. in the underlying population). This is equivalent to saying that the mean outcome difference between the two treatments is **zero** in the population. Under this "null" assumption, we then compute how likely or probable is it that we would get the observed mean difference that we actually observed in our sample or a difference more extreme. This probability is called the **p value**. If the null hypothesis is really true (mean difference is zero in the population), then the mean difference in our sample (or any sample from this population) should usually be small and its corresponding p value should be large. If the null hypothesis is not true, and one treatment really is better than the other in general, then the mean difference in our sample should be large and its corresponding p value should be small. If the p value is small enough, we conclude that the "null" assumption (no true mean difference in the underlying population) is no longer tenable and we "reject" the null hypothesis, thus proving (by contradiction) that one treatment is better, on average, than the other, in general.

This is a very convoluted approach to data analysis since it is not direct proof but proof by contradiction.

Example: HBA1c data (same data as in confidence interval example above).

We know that $\bar{d}$, the mean difference, has a Gaussian distribution with standard error given by $SE_d = \sqrt{}SE_1^2 + SE_2^2$. For the HBA1c data, d = 0.34%, favoring Liraglutide, and the corresponding SE_d = 0.093%.

Under the null hypothesis, $\bar{d}$ should come from a population where the mean differences δ=0. ($\delta=\mu_1-\mu_2$) Since 0.093% is the SE for d, the d we observed, is, under the null hypothesis

Z_{obs} = $(\bar{d} - \delta)/ SE$ = (0.34 – 0) / 0.093 = 3.82 - standard errors from δ=zero.

That is, our observed mean difference of 0.34% HBA1c units is 3.82 SEs from the null value of zero.

According to the Gaussian table, the proportion of samples with an observed Z of 3.82 or greater under the null hypothesis is 0.00008 or 0.008%. The value 0.00008 is the **one sided p value.** By definition, the **two sided p value** is the probability of having an observed Z (Z_{obs}) larger than 3.82 or less than –3.82. This two sided p value is 2 x 0.00008 = 0.00016.

The general convention is to report two sided p values. P values are assumed to be two sided if they are not labeled as one sided or two sided.

Since the p value is small, we "reject" the null hypothesis in this example. That is, we conclude that this observed 0.34% mean difference in our study (or a mean difference more extreme) would only have a very small chance of being observed if there was no true treatment difference in the population. Therefore, by contradiction, the true mean difference can't be zero and the population mean for Liraglutide must be better than the population mean for Sitaglipin. Traditionally, we reject null hypotheses if the two sided p value is less than α=0.05 or the one sided p value is less than α/2=0.025. The "criterion" p value that is the rejection standard (ie. the 0.05) is called the **alpha** level. If α is the two sided alpha level, α/2 is the corresponding one sided alpha level. The value α is the probability that we have made a "false positive" or "type I" mistake by rejecting the null when the null is true (ie saying there is a real difference when there really isn't one). Since this is usually a bad mistake to make, the value of α is usually kept small. Note that the p value comes from and is a function of the data. The alpha level criterion is set in advance by the investigator.

Non inferiority (Equivalence) hypothesis testing

Say that the goal of our study was not to prove that Liraglutin (treatment A) was better than Sitaglipin (treatment B), but to determine if the two treatments were "equivalent", at least on average, or at least that one treatment was not worse than the other. Assume that experts have operationally defined "equivalent" to be "a mean difference no more than δ=0.40%. That is, we are not requiring the two means to be absolutely identical, but only within 0.40 HBA1c units of each other in order to declare "equivalence". In particular, if treatment "B" is an old drug and treatment "A" is a new drug, to show "non inferiority" we need to show that the performance of A is within 0.40% of B in the population. That is:

mean under B minus mean under A < = 0.40% = δ in the population
to show non inferiority

In ordinary hypothesis testing, we assume (the null hypothesis) that the two groups are the same (on average) and test to see if they are different. In non inferiority hypothesis testing, we instead assume that the two groups are different (non equivalent), on average, by a specified "null" amount δ in the population and need to prove that they are equivalent. That is, under non inferiority testing, **the null hypothesis is that the mean difference is δ or more** versus the alternative that the difference is less than δ.

So, under non inferiority testing, we need to see how far our observed $\bar{d}$ is from δ where δ is NOT zero. For example, if δ=**0.40%**, the observed test Z statistic is

Z_{eq} = (0.34 – 0.40) / 0.093 = -0.643. The 0.34 mean is only 0.643 SEs distant from 0.40

One sided p value = 0.26 (area below Z= -0.643).

So, even though we rejected the null hypothesis of no true mean difference earlier, here we cannot reject the null hypothesis of a difference larger than δ in favor of equivalence. That is, we have not **proved** equivalence to within 0.40 **in general** even though the observed mean difference is 0.34, less than 0.40.

Non inferiority test, $Z_{eq} = (\bar{d} - \delta)/ SE_d$

The usual hypothesis testing uses the same formula with δ=0. Since this can be very confusing, many investigators prefer to report the confidence interval. In this example, the 95% confidence interval for the true mean difference of (0.16%, 0.52%) is consistent with both of these findings as it excludes zero and includes 0.30%.

Confidence intervals vs hypothesis testing

It is often easier to examine the confidence interval in order to decide whether to reject or fail to reject null hypotheses, even though one does not need to make any null hypothesis assumptions in order to compute a confidence interval.

In general, if the confidence interval **contains** the null value, the corresponding null hypothesis is NOT rejected. If the confidence interval **excludes** the null value, the corresponding null hypothesis is rejected. Table A below illustrates this concept.

In table A below, those confidence intervals that exclude zero are equivalent to finding statistical significance (i.e. rejecting the null hypothesis that the true difference is zero). Those confidence intervals that are completely between –D and +D demonstrate equivalence (non equivalence is rejected). Those intervals that contain –D or +D but have limits below –D or above +D are uncertain with regard to equivalence (null hypothesis of non equivalence is not rejected).

The confidence level is related to the hypothesis testing p value. If the 95% confidence interval contains the null value, the corresponding two sided p value will be larger than 5%=0.05. If the 95% confidence interval excludes the null value, the corresponding two sided p value will be less than 5%=0.05. In general, if a 100(1-α)% confidence interval excludes the null value, the corresponding two sided p value will be less than "α".

TABLE A

Any confidence interval (95 % CIs intervals between the brackets in each of the examples) that does not overlap zero is statistically different from zero. Only intervals between the prespecified range of equivalence

- D to + D present equivalence.

```
Study         Statistical              equivalence
(1-8)         significance              demonstrated
              demonstrated

1. Yes ------------------------------------------------------------------------------------------- < not equivalent >
2. Yes ------------------------------------------------------------------------------< uncertain >-------------------
3. Yes ------------------------------------------------------------------< equivalent >---------------------------------
4. No ---------------------------------------------------< equivalent >------------------------------------------------
5. Yes ----------------------------------< equivalent >---------------------------------------------------------------
6. Yes --------------------< uncertain>-----------------------------------------------------------------------------
7. Yes -< not equivalent >---------------------------------------------------------------------------------------------
8. No ---------<___________________________uncertain________________________________>------
                              !                                          !
                             -D              O                          +D
                                        true difference
```

Ref: Statistics Applied to Clinical Trials- Cleophas, Zwinderman, Cleopahas 2000 Kluwer Academic Pub Page 35

Paired mean comparison

Paired tests are similar to unpaired tests except the standard error is computed in a different way.

Example: serum cholesterol in mmol/L – before & after treatment

Subject	baseline	4 weeks	difference
1	9.0	6.5	2.5
2	7.1	6. 3	0.8
3	6.9	5.9	1.0
4	6.9	4.9	2.0
5	5.9	4.0	1.9
6	5.4	4.9	0.5
Mean	6.87	5.42	1.45
SD	1.24	0.97	0.79
SE	0.51	0.40	0.32

Mean difference = $\bar{d}$ = 1.45 mmol/L

$SE_d = 0.79/\sqrt{6} = 0.32$ mmol/L , df = 6-1=5

$t_{0.975}$ = 2.571, 95% CI: 1.45 ± 2.571(0.32) or (0.62 mmol/L,2.28 mmol/L)

t obs = $\bar{d}/SE_d$ = 1.45/ 0.32 = 4.49, p value < 0.001

One cannot compute the SE_d using the “Pythagorean” rule for paired comparisons.

Prediction intervals and confidence intervals
They are not the same thing

Don't confuse prediction intervals with confidence intervals.

Standard deviation = SD = measure of an **individuals variability** (about the mean) for continuous data. SDs are usually only quoted for **continuous** data.

But, for **any** statistic and any kind of data (a mean, an OR, a RR, a correlation coefficient…)

Standard error = SE = measure of a **summary statistic's variability** from sample to sample (= from study to study). For a mean the SE is usually denoted SEM (standard error of the mean).

Prediction intervals make a statement about the limits for **individual readings** and are computed from **standard deviations** and a very strong **assumption** that the data follow a Gaussian ("Normal") distribution.

Example: Serum inhibin B in fertile adult men has a mean of 255 pg/ml with a standard deviation of 59 pg/ml. The 95% **prediction** interval is about 255 +/- 1.96(59) or (139 pg/ml, 371 pg/ml). That is, about 95% of the fertile adult men should have serum inhibin B levels within these bounds if the data are well modeled by a Gaussian and if the 255 and 59 are accurate.

Confidence intervals are bounds for **group** summary statistics (means, mean difference, proportions, differences in proportions, risk ratios, odds ratios, slopes, correlations, difference in correlations …), not individuals. Confidence intervals are computed from standard errors, not standard deviations. By the CLT, summary stats usually follow a Gaussian.

For one mean: $\text{SEM} = \text{SD}/\sqrt{n}$,
For the difference between two means: $\text{SE}_d = \sqrt{}\text{SEM}_1^2 + \text{SEM}_2^2$

Imagine that we are interested, in general, in the mean difference in Serum inhibin B between fertile men and men with testicular failure. The sample mean inhibin B in n=8 fertile men is 255 pg/ml and the SD is 59 pg/ml. The sample mean inhibin B in n=16 men with testicular failure is 75 pg/ml and the SD is about 23 pg/ml. The sample mean difference is 255 pg/ml – 75 pg/ml = 180 pg/ml. The SE_d = 21.6 pg/ml

A 95% **confidence** interval for this sample mean difference is
180 pg/ml +/- 1.96 (21.6) pg/ml or (138 pg/ml, 222 pg/ml)

One can imagine taking a census of all fertile men and all men with testicular failure (who are like the men in our sample) and determining the "true" population mean difference in inhibin B for all fertile versus testicular failure men. In practical terms however, we never carry out such a census. In our sample of 16+8= 24 men, our **sample** mean difference in this particular study is 180 pg/ml. When we compute a 95% confidence interval, we are following a process that, 95% of the time, gives an interval that contains the true (unknown) population value. So, we don't know the true population value for certain, but the 95% confidence interval gives a qualitative assessment of how much our sample value may be a good or bad estimate of the true population value.

One can also think of a confidence interval for some sample statistic (such as a mean difference) as indicating how much this sample statistic might vary from study to study if all the studies were done the same way under identical conditions. Obviously, sample statistics tend to vary about the true population value from sample to sample.

Computing confidence intervals and hypothesis test p values

Confidence intervals are of the form

Sample Statistic +/- ($Z_{\text{percentile}}$) (Standard error)

Lower bound = Sample Statistic - ($Z_{\text{percentile}}$) (Standard error)
Upper bound = Sample Statistic + ($Z_{\text{percentile}}$) (Standard error)

The table below shows various sample statistics and their corresponding standard error.

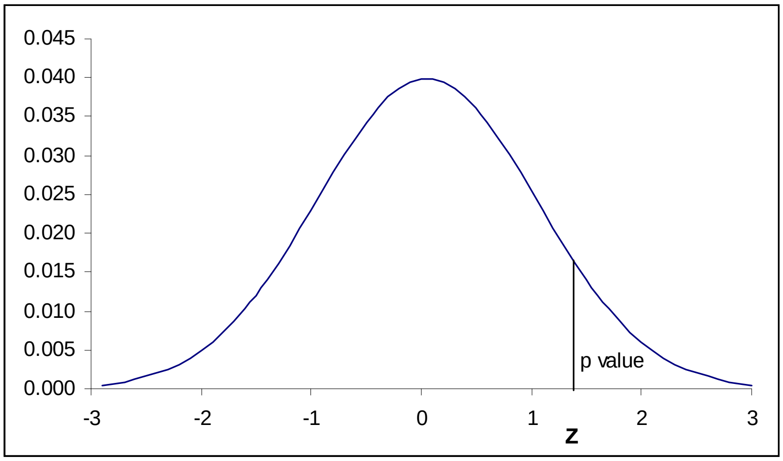
Sample Statistic	**Symbol**	**Standard error (SE)**
Mean	$\overline{Y}$	$S/\sqrt{n} = \sqrt{S^2/n}$ = SEM
Mean difference	$\overline{Y}_1 - \overline{Y}_2 = \overline{d}$	$\sqrt{S_1^2/n_1 + S_2^2/n_2} = SE_d$
Proportion	P	$\sqrt{P(1-P)/n}$
Proportion difference	$P_1 - P_2$	$\sqrt{P_1(1-P_1)/n_1 + P_2(1-P_2)/n_2}$
Log odds	$\log_e(P/(1-P)$	$\sqrt{1/nP + 1/n(1-P)}$
Log odds ratio*	$\log_e OR$	$\sqrt{[1/a + 1/b + 1/c + 1/d]}$
Log risk ratio*	$\log_e RR$	$\sqrt{[1/a - 1/(a+c) + 1/b - 1/(b+d)]}$
Correlation coefficient**	r, $z=\frac{1}{2}\log_e[(1+r)/(1-r)]$	$SE(z)=1/\sqrt{(n-3)}$
Slope (rate)	b	$S_{\text{error}} / S_x\sqrt{(n-1)}$
Hazard (survival)	h	$h/\sqrt{\text{num dead}}$

Any sample statistic has a (sometimes complicated) formula for its standard error that depends on the root sample size, $\sqrt{n}$. Note that, for the OR and RR we first compute the confidence bounds for log(OR) or log(RR) and then take the antilog of each bound.

Hypothesis test statistics (Z_{obs}) are of the form

Z_{obs} = ((Sample Statistic – null value) / Standard error)

where zero (0) is the usual null value. Once this test statistic has been computed, one uses



the appropriate table (such as the Gaussian table) to look up the corresponding p value. (We do not explain here how to actually compute a p value without a table). The p value is the tabled area or probability of getting a test statistic the same as **or more extreme** than the computed Z_{obs} value.

Handy guide to statistical hypothesis testing and power

Hypothesis testing always involves a comparison (to something) and produces a p value, a probability statement that allows "general" conclusions to be drawn about a comparison hypothesis based on data and statistics from a sample.

Hypothesis testing involves a "null" hypothesis – An **assumption** that "nothing is going on" in the underlying population from which the sample data is taken.

Type of sample statistic/comparison	usual population **null** hypothesis
Comparing two means	true pop mean difference is zero
Comparing two proportions	true pop difference between proportions is zero
Comparing two medians	true pop median difference is zero
Odds ratio (comparing odds)	true pop odds ratio is one
Risk ratio = relative risk (comparing risks)	true pop risk ratio is one
Correlation coefficient (compare to zero)	true pop correlation coefficient is zero
Slope= rate of change=regression coeff	true pop slope is zero
Comparing Survival curves	true difference in survival is zero at all times

b. A "p value". This is a probability. However, it is **not** the probability of the null hypothesis given the data. It is the probability of observing the data and corresponding statistic (for example, the mean difference) reported in a given study (or something more extreme) **given that the null hypothesis is true in the underlying population.** Every p values has a null hypothesis that goes with it.

Example: Mean Serum inhibin B concentration is 255 +/- 59 pg/ml (mean +/- SD) in fertile men (n=8) and is 75 +/- 46 pg/ml in men with primary testicular failure (n=16). The mean difference is 255 – 75 = 180 pg/ml. The p value is 0.0001. Taken literally, this p value says that, **if,** in general, in the populations of men who are fertile and men who have testicular failure, the true mean difference (δ) is zero pg/ml, then, we should see a sample mean difference of $\overline{d}$ =180 pg/ml (or greater) only once in every 10,000 studies of this type. (1/10,000 = 0.0001.) Since we in fact did see an average difference of 180 pg/ml in our particular study, we reject the idea that, in general, the mean difference is zero and conclude that, in general, the mean is higher in fertile men compared to men with testicular failure.

(Often, it is easier and more comprehensible to report a confidence interval. One does not need to worry about null hypotheses with confidence intervals).

Testing guide / nomenclature

“δ= Delta” = The true difference or size of the “effect” in the population. For example, delta could be the true population difference between two means or between two proportions. Under the null hypothesis, delta is zero. When “something is going on”, or there is an “effect”, delta is not zero. Delta is sometimes also called the “effect size”. Clinically, one often needs to define the smallest non zero delta that is of any clinical/medical value.

α=Alpha probability = type I error = false “positive” (usually set to alpha = 0.05)
Probability of “rejecting” the null hypothesis when there really is no effect or difference (i.e when the null hypothesis really is true). This is the “criterion” value that we use for deciding if a given p value is “statistically significant” or if a given difference or correlation is significant “beyond chance”. The p value has to be smaller than alpha in order to declare statistical significance.

β= Beta probability = type II error = ‘false “negative”
This is the probability of <u>not</u> rejecting (getting a large p value) when, in fact, there is a real difference (something is going on and delta is not zero).

Power = 1 – Beta = probability of getting a p value less than alpha (i.e declaring “significance”, when, in fact, there really is a non zero delta.

We want small alpha levels and high power.

Check out your intuition. All else held equal:
As delta gets larger, power gets (larger or smaller)
As the sample size gets larger, power gets (larger or smaller)
As alpha gets larger, power gets (larger or smaller)
As patient heterogeneity gets larger (SD gets larger), power gets (larger or smaller)

Statistic/type of comparison	**test/analysis procedure**
Mean comparison-unpaired	t test (2 groups), analysis of variance (ANOVA-3+ groups)
Mean comparison-paired	paired t test, repeated measures ANOVA
Median comparison-unpaired	Wilcoxon rank sum test (2 grp), Kruskal Wallis test*
Median comparison-paired	Wilcoxon signed rank test on differences*, Friedman test*
Proportion comparison-unpaired	chi-square test (or Fishers test)
Proportion comparison-paired	McNemar’s chi-square test
Odds ratio	chi-square test
Risk ratio	chi-square test
Correlation, slope	regression, t statistic
Survival curves, hazard rates	log rank chi-square test

* non parametric – Gaussian distribution theory is not used to get the p value

Parametric versus non parametric p values and confidence intervals

When the distribution of continuous data does not follow the Gaussian (normal bell curve) distribution (or is at least close to it), there is a concern that p values or confidence intervals computed using the Gaussian table (or t table) are not correct, particularly when the sample sizes are small. Therefore, ways have been developed using the ranks of the data values to compute p values and confidence intervals for comparison statistics **without** the Gaussian assumption.

The methods using the Gaussian assumption are often called "parametric" methods since the data distribution is "parameterized" by the Gaussian. If a p value or confidence interval is computed using non Gaussian rank methods (that is, with no Gaussian distribution assumption), this is called a "non parametric" computational method.

Parametric methods		Non parametric methods	
Comparison	method	Comparison	method
comparing two means (2 groups)	t test	comparing two medians (2 groups)	Wilcoxon rank sum test (=Mann-Whitney U test)
paired mean comparison	paired t test	median difference	Wilcoxon signed rank test
comparing several means (3+ groups)	ANOVA	comparing several medians (3+ groups)	Kruskal-Wallis test
Correlation	Pearson correlation	Correlation	Spearman rank correlation

The methods for computing p values and confidence intervals when comparing proportions, odds & odds ratios, risks and risk ratios, hazards and survival curves are generally all non parametric.

selected t percentiles

percentile	85th	90th	95th	97.5th	99.5th	
conf level	70%	80%	90%	95%	99th%	
df = degrees of freedom = sample size minus num unknown parameters						
2	1.386	1.886	2.920	4.303	9.925	
3	1.250	1.638	2.353	3.182	5.841	
4	1.190	1.533	2.132	2.776	4.604	
5	1.156	1.476	2.015	2.571	4.032	
6	1.134	1.440	1.943	2.447	3.707	
7	1.119	1.415	1.895	2.365	3.499	
8	1.108	1.397	1.860	2.306	3.355	
9	1.100	1.383	1.833	2.262	3.250	
10	1.093	1.372	1.812	2.228	3.169	
11	1.088	1.363	1.796	2.201	3.106	
12	1.083	1.356	1.782	2.179	3.055	
13	1.079	1.350	1.771	2.160	3.012	
14	1.076	1.345	1.761	2.145	2.977	
15	1.074	1.341	1.753	2.131	2.947	
20	1.064	1.325	1.725	2.086	2.845	
25	1.058	1.316	1.708	2.060	2.787	
30	1.055	1.310	1.697	2.042	2.750	
35	1.052	1.306	1.690	2.030	2.724	
40	1.050	1.303	1.684	2.021	2.704	
45	1.049	1.301	1.679	2.014	2.690	
50	1.047	1.299	1.676	2.009	2.678	
55	1.046	1.297	1.673	2.004	2.668	
60	1.045	1.296	1.671	2.000	2.660	
65	1.045	1.295	1.669	1.997	2.654	
70	1.044	1.294	1.667	1.994	2.648	
75	1.044	1.293	1.665	1.992	2.643	
80	1.043	1.292	1.664	1.990	2.639	
85	1.043	1.292	1.663	1.988	2.635	
90	1.042	1.291	1.662	1.987	2.632	
95	1.042	1.291	1.661	1.985	2.629	
100	1.042	1.290	1.660	1.984	2.626	
10000	1.036	1.282	1.645	1.960	2.576	<- Gaussian

Section V

Sample size and power

Multiple testing

V- Sample size calculations and power

Sample size for precision / confidence intervals

One basis for computing a sample size is to compute the sample size needed to make a confidence interval of a certain desired width. That is, the sample size is determined in order to achieve a given **precision** needed when estimating the population parameter of interest.

Example:

The Public Health officer wants to sample adult immigrants from Vietnam in order to estimate π, the population prevalence of TB. If P is the proportion with TB in a sample of "n" immigrants, the standard error for P is given by SE = $\sqrt{}$P(1-P)/n. The 95% confidence bounds for the true π, the population TB prevalence, is P +/- 1.96$\sqrt{}$P(1-P)/n Therefore, if we want to estimate prevalence to within +/- 6% of its true value (with 95% confidence),

we need 0.06 = 1.96 $\sqrt{}$P(1-P)/n. Solving for n, we find

$$n= (1.96)^2 P(1-P) /(0.06^2) = 3.84 P(1-P)/ 0.0036.$$

So we now also need to guess at the true value of P. If the true π is 0.15 (ie. 15% have TB), then n = 3.84 0.15(0.85)/ 0.0036 = 136. The largest n is obtained when P=0.5, in which case n = (3.84 x 0.25) / 0.0036 = 267. This is the most conservative value to use. (Conservative rule of thumb: if you want precision +/- w, sample size is less than $1/w^2$.)

We can do similar calculations for any summary statistic, including correlation and regression coefficients and risk and odds ratios.

Power and sample size

More often, we compute the sample size in order to achieve a given statistical **power.** (Usually 80% power or more). Power is a concept under hypothesis testing. As shown below, power is defined as the probability of getting a statistically significant result (ie getting $p < \alpha$) when the null hypothesis is false, that is, when there really is some non zero difference (δ) or, more generally, some non null association. If β is the probability of a "false negative" – that is, of failing to reject the null hypothesis when the null is false, **power is defined as 1-β.**

Hypothesis testing decision table

	Null true (no difference)	Null False (difference)
Test: Don't reject	1-α (correct)	β (error)
Test: Reject	α (error)	(1-β) correct=**power**

Power is computed via a Z score by

$$\mathbf{Z_{power} = Z_{obs} - Z_{\alpha}}$$

where Z_{power} is the Z percentile corresponding to the power, Z_{obs} is the hypothesis testing Z statistic as before and Z_{α} is the Z percentile corresponding to the alpha level. In the case of comparing two means with **assumed** difference d=δ, for the usual α=0.05, Z_{α} = 1.96 and Z_{obs}=d/SE_d, the power is given by

$$\mathbf{Z_{power} = (\bar{d}/SE_d) - 1.96}$$

Example: Mean HBA1c change difference under two treatments
(Same example as above with much smaller sample size)

HBA1c change

Treatment	n	Mean change	SD	SEM
Liraglutide	5	-1.24	0.99	0.443
Sitaglipin	4	-0.90	0.98	0.488

For this data, $\bar{d}$= 0.34% $SE_d = \sqrt{0.443^2 + 0.488^2} = 0.659$
Z_{obs} = 0.34/0.659 = 0.516, two sided p value = 0.622.

Since p = 0.622, we say "not statistically significant" using the α=0.05 criterion. This does NOT mean that we have shown Liraglutide to be equivalent to Sitaglipin. We ask "what is our power?".

In this example, Z_{power} = 0.516– 1.96 = -1.44. Using the Gaussian table, Z=-1.44 is about the 7th percentile. So our power here is only about 7%.

This means that, even if the true δ in the population is not zero and favors Liraglutide, so that, on average, Liraglutide really is better than Sitaglipin, using the p < 0.05= α criterion and these sample sizes, we would only get statistical significance about 7% of the time. We "didn't have a chance" of getting statistical significance, even if Liraglutide really is better than Siraglipin in general.

Low power implies that we failed to find significance because we did not collect enough data, not that the null hypothesis is correct. Only if we fail to reject and have high power can we conclude that the null hypothesis is proven.

It is incumbent upon investigators to compute power for the smallest true mean difference (δ) that is clinically worthwhile when planning a study. This will help avoid interpreting a "non significant" result as a "negative" result. This is best done during the planning stage if the intent it to have a definitive results.

In general, if a result is not statistically significant, we want at least **80%** power in order to conclude that the null hypothesis is actually true.

It is wrong to conclude that a non significant result is a negative result. In order to really affirm that we have a negative result, we also must show that we have high power for the smallest true difference of interest.

Obviously, the better course of action is to estimate the sample size needed to obtain at least 80% power (or some other high power level) before we start a study. This is why many research sponsors (such as the NIH) will not fund a study unless there is high power (say, 80% or 90% power) for the primary outcome.

Solving the above formula for "n", tables and computer software are available that give the required sample size for a given power level and the power for a given sample size.

In the case of comparing two means between two groups with the same sample size of "n" per group,

$$n = 2\ (Z_{power} + Z_{\alpha})^2\ (\sigma/\delta)^2.$$

For 80% power, $Z_{power} = 0.842$, for $\alpha=0.05$, $Z_{\alpha} = 1.96$ so, in this case

$$n=2(0.842+1.96)^2\ (\sigma/\delta)^2 \text{ or } n = 15.7(\sigma/\delta)^2 \approx (4\sigma/\delta)^2 \approx \mathbf{(range/\delta)^2}$$

Summary - Power gets larger as:

a. The true difference (δ) gets larger
b. The sample size gets larger
c. The α level gets larger (ie. less strict significance criterion)
d. The patient heterogeneity (σ) gets smaller

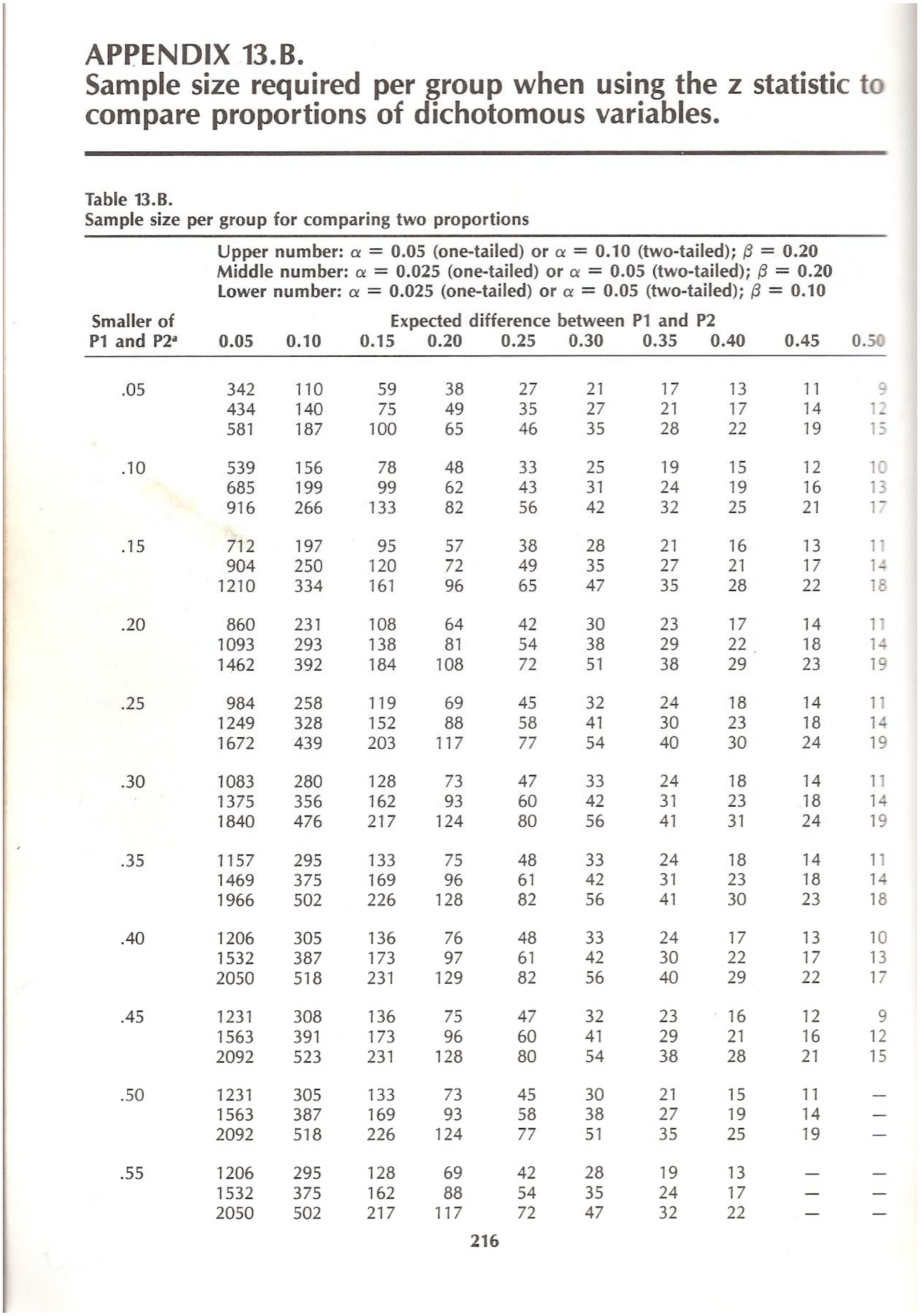
Generally, we set power = $1-\beta = 0.80$ and $\alpha = 0.05$. Therefore, in order to determine the sample size, you must have at least a crude estimate of δ and σ. We generally set δ to be the smallest difference that is clinically important. We estimate σ from past data on similar patients (often from the literature) or get some pilot data on a few new patients or use the best clinical judgment. A crude rule based on Gaussian theory is to set σ to the range/4 after removing any outliers.

We can do power/sample size computation for any summary statistic, not just mean differences. The methods are similar to the above. When comparing proportions (and their corresponding odds ratios and/or risk ratios), σ is a function of the proportions so separate prior knowledge of σ is not needed. However, one needs to specify the two proportions, not just the difference δ between them.

Sample size per group for comparing two means
from two independent groups

mean diff = smallest mean difference of interest   alpha = 0.05, two sided

mean diff/SD= δ/σ	**70% power**	**80% power**	**90% power**
0.10	1234	1570	2102
0.15	549	698	934
0.20	309	392	525
0.25	198	251	336
0.30	137	174	234
0.35	101	128	172
0.40	77	98	131
0.45	61	78	104
0.50	49	63	84
0.55	41	52	69
0.60	34	44	58
0.65	29	37	50
0.70	25	32	43
0.75	22	28	37
0.80	19	25	33
0.85	17	22	29
0.90	15	19	26
0.95	14	17	23
1.00	12	16	21
1.05	11	14	19
1.10	10	13	17
1.15	9	12	16
1.20	9	11	15
1.25	8	10	13
1.30	7	9	12
1.35	7	9	12
1.40	6	8	11
1.45	6	7	10
1.50	5	7	9



APPENDIX 13.B.
Sample size required per group when using the z statistic to compare proportions of dichotomous variables.

Table 13.B.
Sample size per group for comparing two proportions

Upper number: $\alpha = 0.05$ (one-tailed) or $\alpha = 0.10$ (two-tailed); $\beta = 0.20$
Middle number: $\alpha = 0.025$ (one-tailed) or $\alpha = 0.05$ (two-tailed); $\beta = 0.20$
Lower number: $\alpha = 0.025$ (one-tailed) or $\alpha = 0.05$ (two-tailed); $\beta = 0.10$

Smaller of	Expected difference between P1 and P2									
P1 and P2[a]	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
.05	342	110	59	38	27	21	17	13	11	9
	434	140	75	49	35	27	21	17	14	12
	581	187	100	65	46	35	28	22	19	15
.10	539	156	78	48	33	25	19	15	12	10
	685	199	99	62	43	31	24	19	16	13
	916	266	133	82	56	42	32	25	21	17
.15	712	197	95	57	38	28	21	16	13	11
	904	250	120	72	49	35	27	21	17	14
	1210	334	161	96	65	47	35	28	22	18
.20	860	231	108	64	42	30	23	17	14	11
	1093	293	138	81	54	38	29	22	18	14
	1462	392	184	108	72	51	38	29	23	19
.25	984	258	119	69	45	32	24	18	14	11
	1249	328	152	88	58	41	30	23	18	14
	1672	439	203	117	77	54	40	30	24	19
.30	1083	280	128	73	47	33	24	18	14	11
	1375	356	162	93	60	42	31	23	18	14
	1840	476	217	124	80	56	41	31	24	19
.35	1157	295	133	75	48	33	24	18	14	11
	1469	375	169	96	61	42	31	23	18	14
	1966	502	226	128	82	56	41	30	23	18
.40	1206	305	136	76	48	33	24	17	13	10
	1532	387	173	97	61	42	30	22	17	13
	2050	518	231	129	82	56	40	29	22	17
.45	1231	308	136	75	47	32	23	16	12	9
	1563	391	173	96	60	41	29	21	16	12
	2092	523	231	128	80	54	38	28	21	15
.50	1231	305	133	73	45	30	21	15	11	—
	1563	387	169	93	58	38	27	19	14	—
	2092	518	226	124	77	51	35	25	19	—
.55	1206	295	128	69	42	28	19	13	—	—
	1532	375	162	88	54	35	24	17	—	—
	2050	502	217	117	72	47	32	22	—	—

Ref – Hulley & Cummings, Designing Clinical Research, 1988, Williams & Wilkins

Sample size checklist

One of the most important questions when planning a study and meeting with a statistician is determining the number of subjects needed to carry out the study and get a definitive answer.

In order to determine a sample size, one must know (or guess at):

Effect size ("delta") - size of treatment effect such as a mean difference, a difference in proportions, a difference in rates, a ratio – ***what is the smallest difference (on average) or ratio that is clinically important?*** Sample size decreases as effect size increases.

Time of comparison – For a time dependent outcome, the time it takes to achieve the effect. For example, we might say that a reduction from 50% to 30% is expected after two years. The sooner the difference specified occurs, the smaller the sample needed.

Follow up time – For time dependent outcomes, the amount of time persons are followed is as important as the number of people. The required sample size can be reduced somewhat without loss of power if persons are followed longer.

Variability – Patient heterogeneity such as the within group standard deviation. Sample size increases as variability increases.

Power – Usually a minimum of 80% power is required (NIH). Sample size increases if a higher power is required.

Alpha level – This is usually set at 0.05 two sided. Sample size decreases if a more liberal (larger) alpha is used. But a larger alpha is a higher false positive error rate.

The above gives the sample size if there were no dropout or loss to follow up

---

In addition, on any prospective study with recruitment over time, one must give:

The percentage who will agree to participate = 100% - refusal percentage
(or sometimes the refusal rate per unit of time)

The accrual rate (per unit time)

The dropout / loss to follow up rate (per unit time)

Hypothesis testing & sample size limitations

Pseudo replication

Most variation is between person, not within person.

A common mistake made with regard to sample size is confusing the number of observations made on each person (or each "experimental unit") with the number of persons in a study.

Example: If m=2 two blood samples are taken on n=10 persons, the "effective" sample size is 10, not 20.

In general, the standard error for any group statistic, such as a mean, is now influenced by two sources of variation, within person variation and between person variation.

Observed observation =
true population parameter + between person variation + within person variation

Under this circumstance, a mean is computed in two stages:

1. Compute a mean for each person from her "m" observations
2. Compute the group mean from the "n" individual person means.

Other summary statistics are often computed similarly.

If there are "m" observations per person and "n" people, the standard error (SE) for the group mean is given by

SE of the mean = SEM = $\surd\sigma_p^2/n + \sigma_e^2/nm$

where σ_p is the standard deviation of the between person variation and σ_e is the standard deviation of the within person variation. In most circumstances, $\sigma_e < \sigma_p$, so, to a reasonable approximation, the SEM is equal to $\sigma_p/\surd n$ as usual. That is, the contribution of σ_e^2/nm to the SEM is negligible, particularly when m is large. When m=1, (one observations per person) one cannot distinguish between σ_p and σ_e.

Statistical significant versus clinical/scientific significance

While it takes substantial effort to compute p values and determine whether effects are “beyond chance”, one must not forget that p value based “statistical significance” is NOT the same as clinical importance. The much higher priority is to first determine whether results are of scientific or clinical importance, not whether they are “statistically significant” (“A difference, in order to be a difference, must make a difference”–Gertrude Stein?).

If you are in charge of research funding, which of the two studies below, I or II, should be given more research funds?

Average drop in weight (kg) after dieting for 3 months

Diet	**mean drop**	**p value**	**95% confidence interval**
I	**0.50**	**< 0.001**	**(0.45, 0.55)**
II	**10.0**	**0.16**	**(-5.0, 25.0)**

***

p value limitations

Reporting p values alone is not enough. The American Statistical Association (ASA) lists 5 limitations.

1. p values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone (ignoring the model).
2. Conclusions should not be based only on whether a p-value passes a specific threshold.
3. Proper inference requires full reporting & transparency.
4. A p-value does not measure the size of an effect or the importance of a result.
5. A p-value alone does not provide a good measure of evidence regarding a model or hypothesis

Multiple hypothesis testing

"If you torture the data long enough, it will eventually confess"

Multiple testing- Multiple analyses:
  Multiple efficacy endpoints /outcomes
  Multiple safety endpoints/outcomes
  Multiple treatment arms and/or doses
  Multiple interim analyses
  Multiple patient subgroups

Example- Exploratory vs confirmatory - protein comparison

750 proteins are compared in two groups. There are statistically significant differences at p < 0.05 in 12 proteins.

Protein name	Atril fib	Atherosclerosis	p value
RAS guanyl-releasing protein 2	33.3%	0.0%	0.0000
Glutathione S-transferase P	38.9%	100.0%	0.0000
Selenium-binding protein 1	22.2%	0.0%	0.0000
Nucleosome assembly protein 1-like 4	16.7%	0.0%	0.0000
Integrin beta;Integrin beta-2	11.1%	50.0%	0.0000
Spectrin alpha chain, non-erythrocytic 1	11.1%	0.0%	0.0000
Pituitary tumor-transforming gene 1 protein-interacting	11.1%	0.0%	0.0000
WW domain-binding protein 2	16.7%	50.0%	0.0000
Syntaxin-4	5.6%	0.0%	0.0006
CD9 antigen	27.8%	50.0%	0.0013
ATP synthase-coupling factor 6, mitochondrial	27.8%	50.0%	0.0013
Flotillin-1	77.8%	100.0%	0.0037
Aconitate hydratase, mitochondrial	38.9%	50.0%	0.1142
Fructose-bisphosphate aldolase C	94.4%	100.0%	0.4402
Alpha-adducin	50.0%	50.0%	1.0000
40S ribosomal protein SA	1.0%	1.0%	1.0000
Abl interactor 1	1.0%	1.0%	1.0000
Bone marrow proteoglycan;Eosinophil granule major basic	1.0%	1.0%	1.0000
Tubulin alpha-4A chain	100.0%	100.0%	1.0000
… (750 proteins total)			

Example - The “disease score” ranges from 2 (good) to 12 (worst).

Scenario A: Due to prior suspicion (prior information), only patients 19 and 47 are measured and both have scores of 12. We report that they are “significantly” ill.

Scenario B: The score is measured on 72 patients. Only patients 19 and 47 have scores of 12. We report that they are “significantly” ill.

Is the amount of “evidence” or “belief” that patients 19 and 47 “really” are very ill (have “true” score of 12) the same in both scenarios? The data for patients 19 and 47 are the same in both scenarios.

Most would agree that, if both patients were retested (confirmation step), and came out with lower scores, this would decrease the belief that there “true” score is 12. If they came out with 12 again, this would increase the belief that the true score is 12.

Example - Imagine two different situations in evaluating a new drug for arthritis compared to aspirin.

A. Only pain (0-10) and swelling (0-10) are measured and both are significantly better at $p < 0.05$ on the new treatment compared to aspirin.

B. Ten different outcomes are measured: pain, swelling, activities of daily living, quality of life, sleep, walking, bending, lifting, grinding, climbing. Of these 10, the investigators only report the two (pain, swelling) that were better on the new treatment after looking at the results for all ten. They fail to report the other eight, which were not significant at $p < 0.05$.

Issue 1 – It is grossly misleading to only publish the results from these two tests and **not** reveal that the other 8 were examined, with “negative” results.

In a **confirmatory** study, must state what comparisons / analyses will be done **in advance**.

Say, instead, the investigators did report all 10.

Issue 2- (Bonferroni) Out of **m** tests, if they are all independent and we use the $p < 0.05$ criterion, we expect, on average, that 0.05 m of them to come out "significant" by chance alone even if none of the **m** really have any true effect in the population.

Num of tests	probability at least one is significant at 0.05
(m)	even if null hypothesis is true for all*
1	0.0500
2	0.0975
3	0.1426
4	0.1855
5	0.2262
10	0.4013
20	0.6415
25	0.7226
50	0.9231

* assuming independence

Usually, the tests are not all independent so results are not this bad, but the above should convey the general issue.

What to do about multiple testing – Holm/Hochberg criterion

Option 1 – Use the nominal α level for each test. That is declare significant if p value $< \alpha$. This can produce too many false positives (type I error $> \alpha$) but keeps power at $1-\beta$.

Option 2 – Use the Bonferroni criterion (not recommended). Declare significant only if $p < \alpha/m$. This is **very** conservative if m is large and therefore can produce too many false negatives (low power) but guarantees that the overall type I error is $\leq \alpha$.

Option 3 – Use the Holm-Hochberg (H-H) criterion / rule.

a. For m significance tests, sort the "m" p values from smallest to largest. Denote the smallest p_1, next smallest p_2, …. $p_{m.}$

b. Declare the i^th ordered p value (p_i) significant **only if $p_i < \alpha/(m+1-i)$**. If, for i=k, $p_k \geq \alpha/(m+1-k)$, then p_k and all subsequent larger ordered p values (p_{k+1}, p_{k+2}, … p_m) are declared non significant.

The H-H rule keeps the overall type I error ≤ α.

The -H rule guarantees that the overall false positive rate will be ≤ α. It is a compromise between options 1 and 2 above.

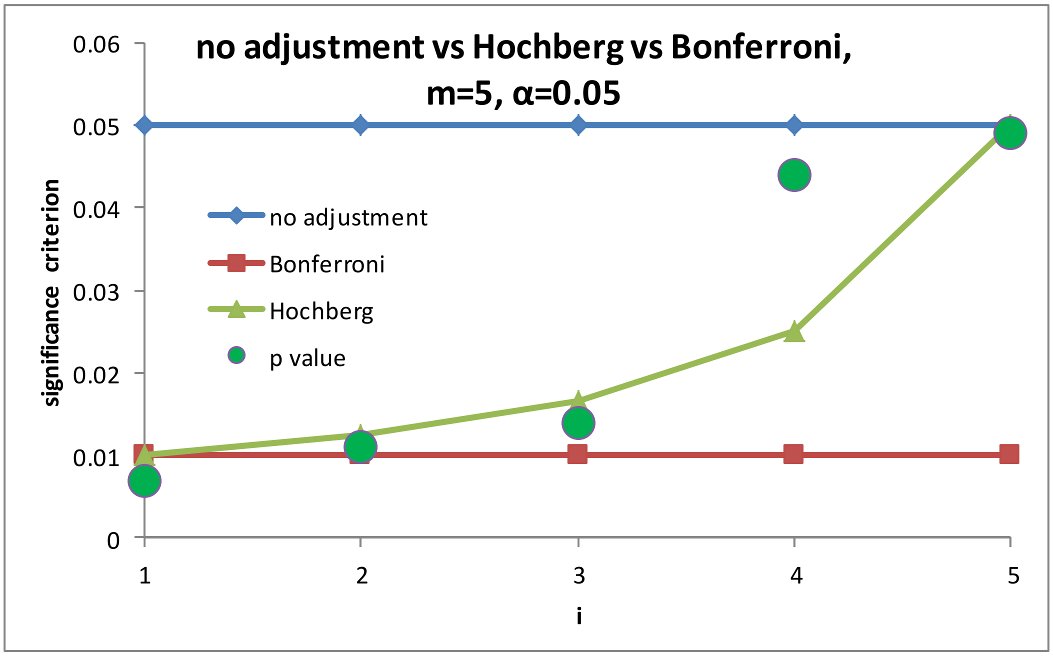
Example for m=5

i	p value	criterion = (i/5)α	criterion if α=0.05
1	p_1-smallest	α/5	0.0100
2	p_2	α/4	0.0125
3	p_3	α/3	0.0167
4	p_4	α/2	0.0250
5	p_5-largest	α	0.0500

Bonferroni significance criterion is $p < \alpha/m=0.05/5=0.01$

There are several other options, most of which we will not study. (See ANOVA section for methods specific to ANOVA).

Later we will study omnibus screening tests such as the F and chi-square tests.



FDR, a more liberal alternative to FWER

If a "family" of "m" hypothesis tests are carried out, the family wise error rate (FWER) is the chance of any "false positive" type I error assuming that the null is true for all m tests.

Rather than control the FWER, it may be preferable to control the number of "positive" tests (not all m tests) that are false positives. This is called controlling the false discovery rate (FDR), a less stringent criterion.

For FDR, the ith ordered p value must be less than **(i/m)α** which is larger than Holm-Hochberg α/(m+i-i) for FWER.

	Declare non sig	Declare sig	**total**
Truth-Null True	U	V	**m_0**
Truth-Null False	T	S	**$m-m_0$**
total	**m-R**	**R**	**m**

FWER = V/m, the probability that V ≥ 1,   FDR = V/R, more liberal

FWER vs FDR , m=5 hypothesis tests, α=0.05

p value	FDR criteria	FWER criteria
p1-smallest	(1/5)α=0.01	α/5=0.01
p2	(2/5)α=0.02	α/4=0.0125
p3	(3/5)α=0.03	α/3=0.0167
p4	(4/5)α=0.04	α/2=0.025
p5-largest	(5/5)α=0.05	A=0.05

Multiple testing and designating primary versus secondary outcomes

When there are “m” outcomes, we must have a stricter alpha level criterion for each outcome in order to control for the overall type I error in all m. This leads to a larger sample size if alpha is smaller.

However, in most studies, not all outcomes or endpoints may be equally important. Therefore it is common to designate the most important outcomes “primary” and let m be the number of primary outcomes. That is, one only controls for the overall type I error rate in the primary outcomes. Since the secondary outcomes are less important, one does not include them in the “family” of m primary outcomes and one is therefore not as concerned if there is a false positive finding among the outcomes designated as secondary.

However, one must designate the primary and secondary outcomes in advance, before the results of the study are known. It is not fair to declare which outcomes are primary and which are secondary based on their p values.

Statistical Analysis Plan outline

Statistical models and methods to answer study questions (Aims)

Conclusions = data + models (assumptions)

Each specific aim needs a stat analysis section.

Sample size and power follows the analysis plan.

Outline:

- Outcomes: denote primary & secondary
- Primary predictors or comparison groups
- Covariates/confounders/effect modifiers
  - Methods for missing data, dropouts
  - Interim analyses (for efficacy, for safety)