Introduction to Biostatistics for Biomedical Research

Frank E Harrell Jr
James C Slaughter
Department of Biostatistics
Vanderbilt University School of Medicine
f.harrell@vanderbilt.edu
james.c.slaughter@vanderbilt.edu

biostat.mc.vanderbilt.edu/ClinStat

Updated February 18, 2008

Contents

Note: Symbols in boxes in the right margin designate page numbers in Altman, Machin, Bryant, and Gardner (numbers given by Ann), section numbers in Katz (numbers given by Knn), or section numbers in Essential Medical Statistics (EMSnn).

Chapter 1

General Overview

K1-3

EMS2

1.1 What is Biostatistics?

- Statistics applied to biomedical problems
- Decision making in the face of uncertainty or variability
- Design and analysis of experiments; detective work in observational studies (in epidemiology, outcomes research, etc.)
- Attempt to remove bias or find alternative explanations to those posited by researchers with vested interests
- Experimental design, measurement, description, statistical graphics, data analysis, inference

1.2 Types of Data Analysis and Inference

- Description: what happened to *past* patients

- Inference from specific (a sample) to general (a population)
  - Hypothesis testing: test a hypothesis about population or long-run effects
  - Estimation: approximate a population or long term average quantity
  - Prediction: predict the responses of other patients *like yours* based on analysis of patterns of responses in your patients

1.3 Types of Measurements by Their Role in the Study

K3

- Response variable (clinical endpoint, final lab measurements, etc.)
- Independent variable (predictor or descriptor variable) — something measured when a patient begins to be studied, before the response; often not controllable by investigator, e.g. sex, weight, height, smoking history
- Adjustment variable (confounder) — a variable not of major interest but one needing accounting for because it explains an apparent effect of a variable of major interest or because it describes heterogeneity in severity of risk factors across patients
- Experimental variable, e.g. the treatment or dose to which a patient is randomized; this is an independent variable under the control of the researcher

Table 1.1: *Common alternatives for describing independent and response variables*

Response variable	Independent variable
Outcome variable	Exposure variable
Dependent variable	Predictor variable
y-variables	x-variable
Case-control group	Risk factor
	Explanatory variable

1.4 Types of Measurements According to Coding

K3

EMS2.2

- Binary: yes/no, present/absent
- Categorical (nominal, polytomous, discrete): more than 2 values that are not necessarily in special order
- Ordinal: a categorical variable whose possible values are in a special order, e.g., by severity of symptom or disease; spacing between categories is not assumed to be useful
- Count: a discrete variable that (in theory) has no upper limit, e.g. the number of ER visits in a day, the number of traffic accidents in a month
- Continuous: a numeric variable having many possible values representing an underlying spectrum
- Continuous variables have the most statistical information (assuming the raw values are used in the data analysis) and are usually the easiest to standardize across hospitals
- Turning continuous variables into categories by using intervals of values is arbitrary and requires more patients to yield the same statistical information (precision or power)
- Errors are not reduced by categorization unless that's the only way to get a subject to answer the question (e.g., income[a]

[a] But note how the Census Bureau tries to maximize the information collected. They first ask for income in dollars. Subjects refusing to answer are asked to choose from among 10 or 20 categories. Those not checking a category are asked to choose from fewer categories.

1.5 Random Variables

- A potential measurement X

- X might mean a blood pressure that will be measured on a randomly chosen US resident

- Once the subject is chosen and the measurement is made, we have a sample value of this variable

- Statistics often uses X to denote a potentially observed value from some population and x for an already-observed value (i.e., a constant)

Chapter 2

Descriptive Statistics and Distributions

EMS3

2.1 Distributions

K4

The *distribution* of a random variable X is a profile of its variability and other tendencies. Depending on the type of X, a distribution is characterized by the following.

- Binary variable: the probability of “yes” or “present” (for a population) or the proportion of same (for a sample).

- k-Category categorical (polytomous, multinomial) variable: the probability that a randomly chosen person in the population will be from category $i, i = 1, \ldots, k$. For a sample, use k proportions or percents.

- Continuous variable: any of the following 4 sets of statistics
  - probability density: value of x is on the x-axis, and the relative likelihood of observing a value “close” to x is on the y-axis. For a sample this yields a histogram.

  - cumulative probability distribution: the y-axis contains the probability of observing $X \leq x$. This is a function that is always rising or staying flat, never decreasing. For a sample it corresponds to a cumulative histogram[a]

  - all of the *quantiles* or *percentiles* of X

  - all of the *moments* of X (mean, variance, skewness, kurtosis, . . .)

  - If the distribution is characterized by one of the above four sets of numbers, the other three sets can be derived from this set

- Knowing the distribution we can make intelligent guesses about future observations from the same series, although unless the distribution really consists of a single point there is a lot of uncertainty in predicting an individual new patient's response. It is less difficult to predict the average response of a group of patients once the distribution is known.

- At the least, a distribution tells you what proportion of patients you would expect to see whose measurement falls in a given interval.

2.1.1 Distribution Shapes

2.2 Descriptive Statistics

K4

EMS3.2-3.3

2.2.1 Categorical Variables

- Proportions of observations in each category
  Note: The mean of a binary variable coded 1/0 is the proportion of ones.

- For variables representing counts (e.g., number of comorbidities), the mean

[a] But this *empirical cumulative distribution function* can be drawn with no grouping of the data, unlike an ordinary histogram.

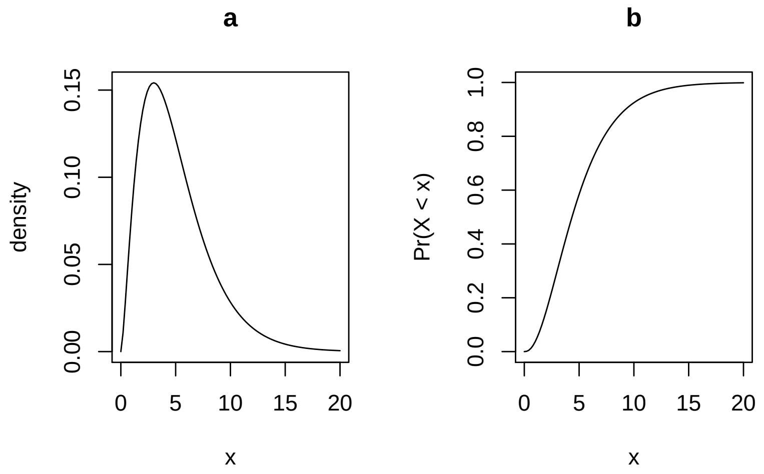


Figure 2.1: *Example probability density (a) and cumulative probability distribution (b)*

is a good summary measure (but not the median)

- Modal (most frequent) category

2.2.2 Continuous Variables

Denote the sample values as $x_1, x_2, \ldots, x_n$

Measures of Location

“Center” of a sample

- Mean: arithmetic average

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

  Population mean μ is the long-run average (let $n \to \infty$ in computing $\bar{x}$)

  - center of mass of the data (balancing point)

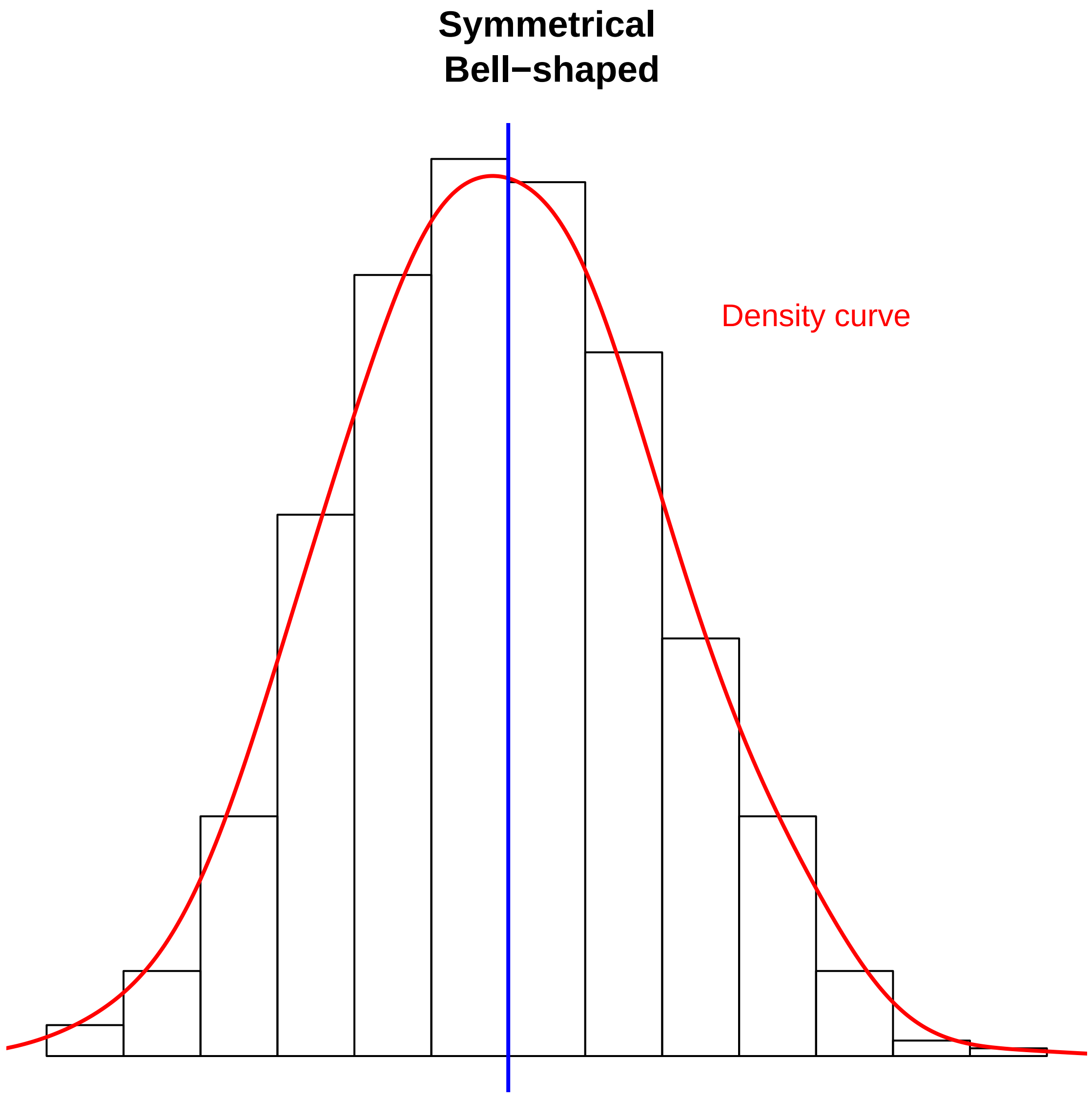


Figure 2.2: *Example of a distribution that is symmetric about the mean (blue line). Sample from a Normal distribution.*

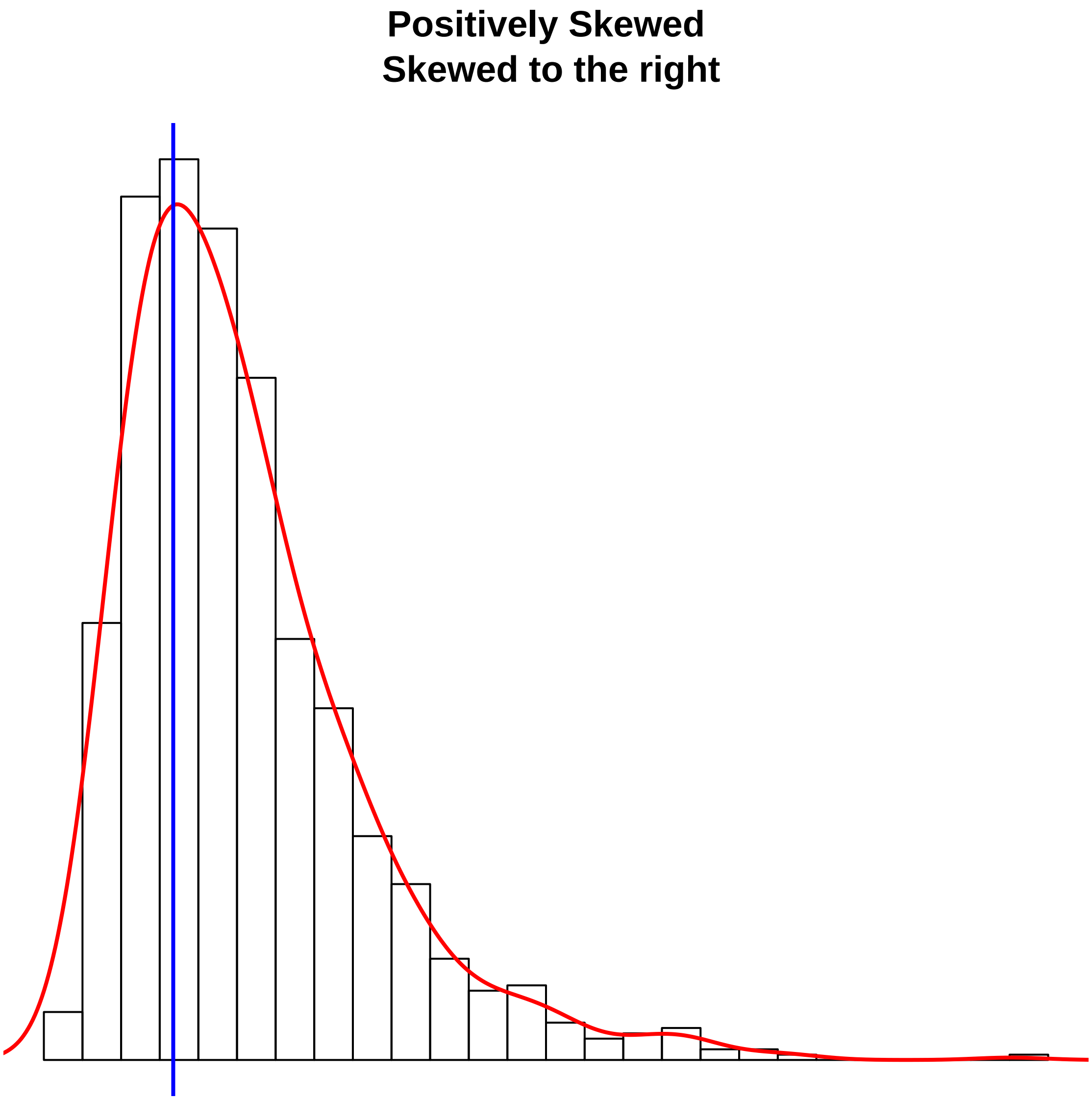


Figure 2.3: *Example of a distribution that is skewed to the right. Sample is from a log-Normal distribution.*

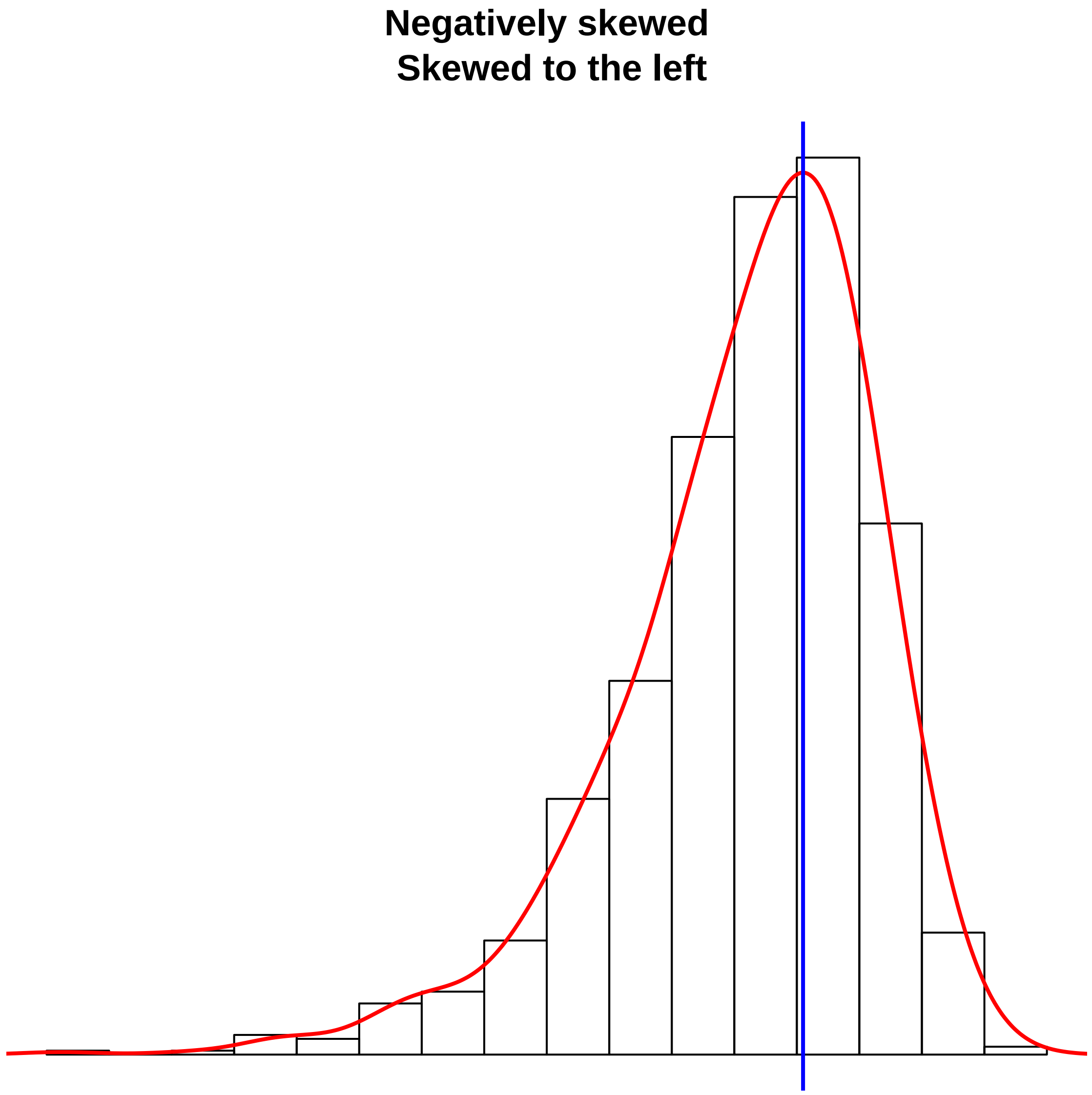


Figure 2.4: *Example of a distribution that is skewed to the left*

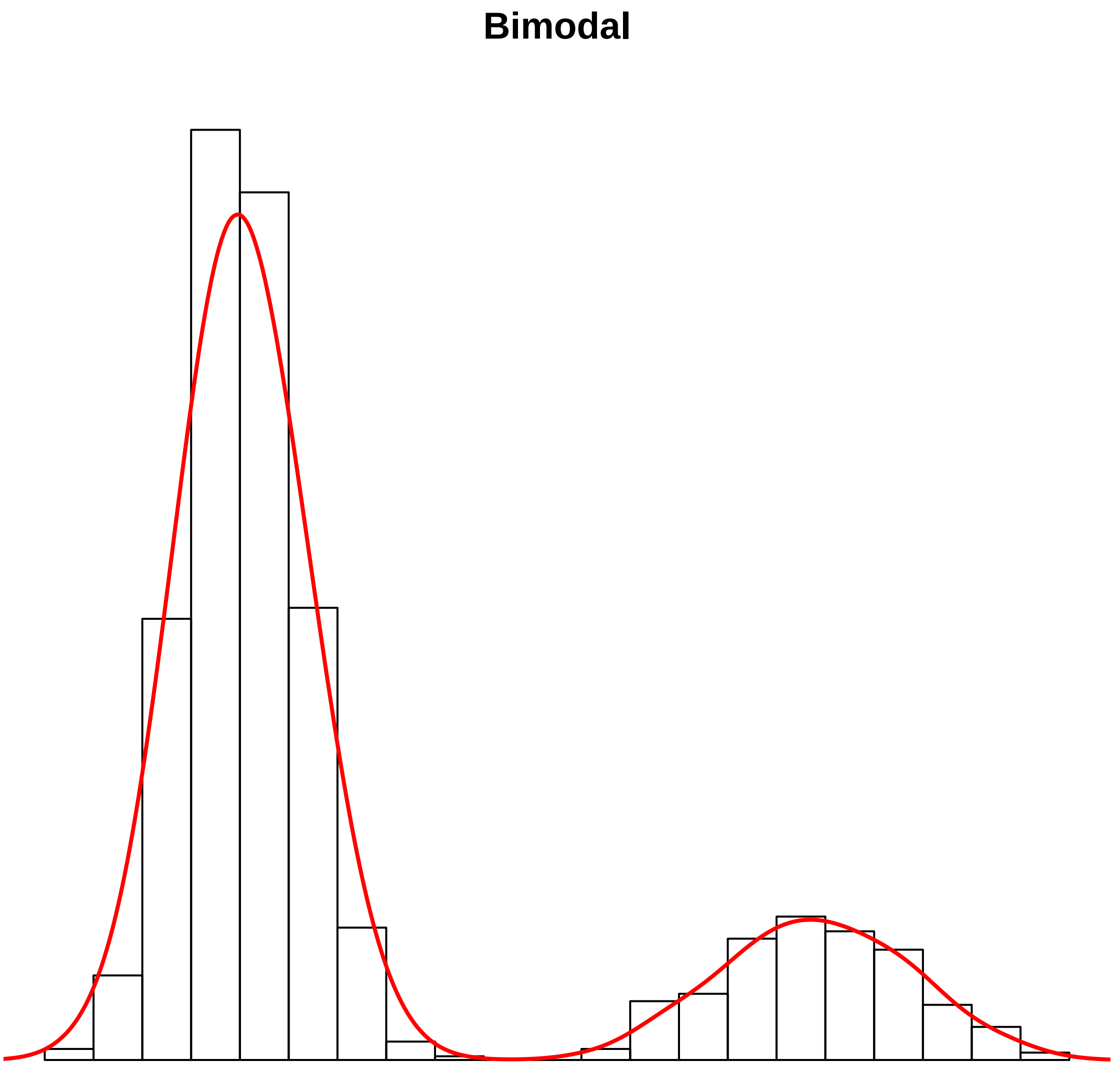


Figure 2.5: *Example of a distribution that is bimodal (has two peaks). Sample is from a mixture of two Normal distributions.*

  - highly influenced by extreme values even if they are highly atypical

- Median: middle sorted value, i.e., value such that $\frac{1}{2}$ of the values are below it and above it
  - always descriptive
  - unaffected by extreme values
  - not a good measure of central tendency when there are heavy ties in the data
  - if there are heavy ties and the distribution is limited or well-behaved, the mean often performs better than the median (e.g., mean number of diseased fingers)

- Geometric mean: hard to interpret and effected by low outliers; better to use median

Quantiles

Quantiles are general statistics that can be used to describe central tendency, spread, symmetry, heavy tailedness, and other quantities.

- Sample median: the 0.5 quantile or 50^{th} percentile
- Quartiles Q_1, Q_2, Q_3: 0.25 0.5 0.75 quantiles or $25^{th}, 50^{th}, 75^{th}$ percentiles
- Quintiles: by 0.2
- In general the pth sample quantile x_p is the value such that a fraction p of the observations fall below that value

- p^{th} population quantile: value x such that the probability that $X \leq x$ is p

Spread or Variability

- Interquartile range: Q_1 to Q_3
  Interval containing $\frac{1}{2}$ of the subjects
  Meaningful for any continuous distribution
- Other quantile intervals
- Variance (for symmetric distributions): averaged squared difference between a randomly chosen observation and the mean of all observations

  $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  The -1 is there to increase our estimate to compensate for our estimating the center of mass from the data instead of knowing the population mean.[b]
- Standard deviation: s — $\surd$ of variance
  - $\surd$ of average squared difference of an observation from the mean
  - can be defined in terms of proportion of sample population within $\pm$ 1 SD of the mean **if the population is normal**
- SD and variance are not useful for very asymmetric data, e.g. “the mean hospital cost was \$10000 $\pm$ \$15000”
- range: not recommended because range $\uparrow$ as $n \uparrow$ and is dominated by a single outlier
- coefficient of variation: not recommended (depends too much on how close the mean is to zero)

[b] $\bar{x}$ is the value of μ such that the sum of squared values about μ is a minimum.

2.3 Graphs

K4.2-4.6

A17-20

2.3.1 Categorical Variables

- pie chart
  - high ink:information ratio
  - optical illusions (perceived area or angle depends on orientation vs. horizon)
  - hard to label categories when many in number
- bar chart
  - high ink:information ratio
  - hard to depict confidence intervals (one sided error bars?)
  - hard to interpret if use subcategories
  - labels hard to read if bars are vertical
- dot chart
  - leads to accurate perception
  - easy to show all labels; no caption needed
  - allows for 3 levels of categorization (see Figure 2.6)
    - multi-panel display for multiple major categorizations
    - lines of dots arranged vertically within panel

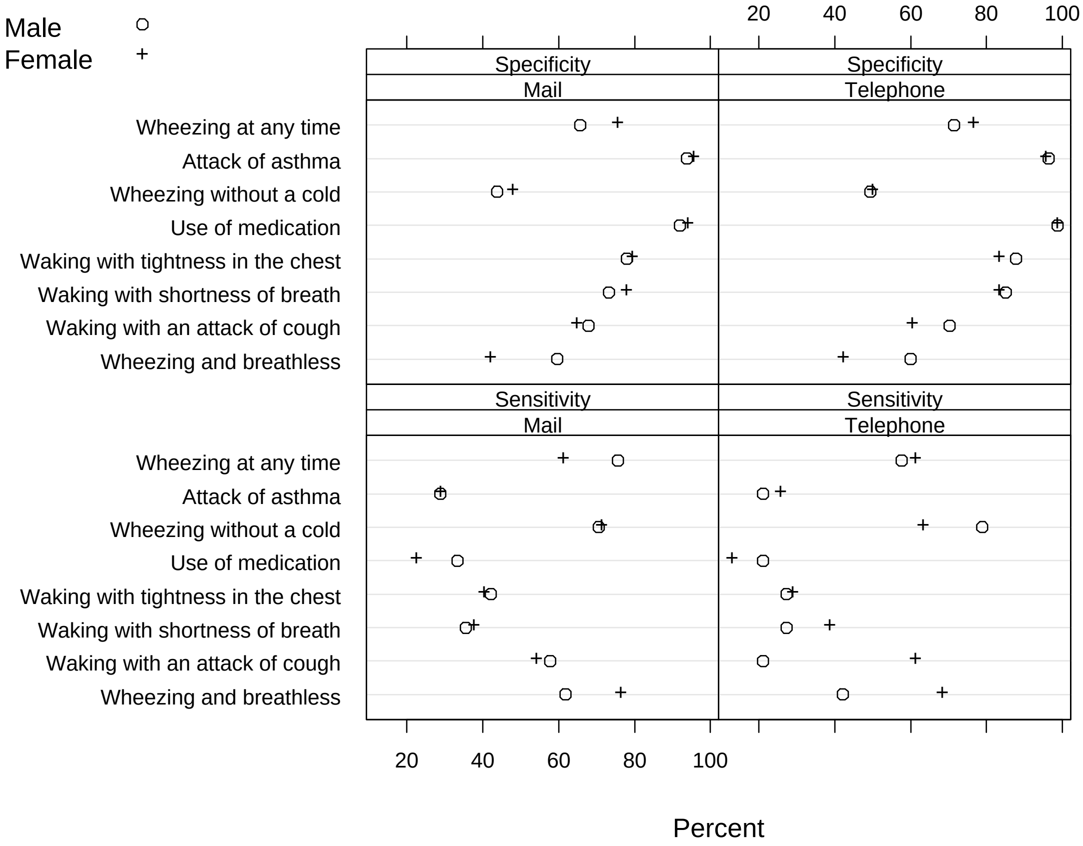


Figure 2.6: *Dot chart*

    - $*$ categories within a single line of dots

  - easy to show 2-sided error bars

- Avoid chartjunk such as dummy dimensions in bar charts, rotated pie charts, use of solid areas when a line suffices

2.3.2 Continuous Variables

Distributions

- histogram showing relative frequencies
  - requires arbitrary binning of data

  - not optimal for comparing multiple distributions

- cumulative distribution function: proportion of values $\leq x$ vs. x (Figure 2.7) Can read all quantiles directly off graph.

- box plot: quartiles plus mean. Good way to compare many groups (Figure 2.8)

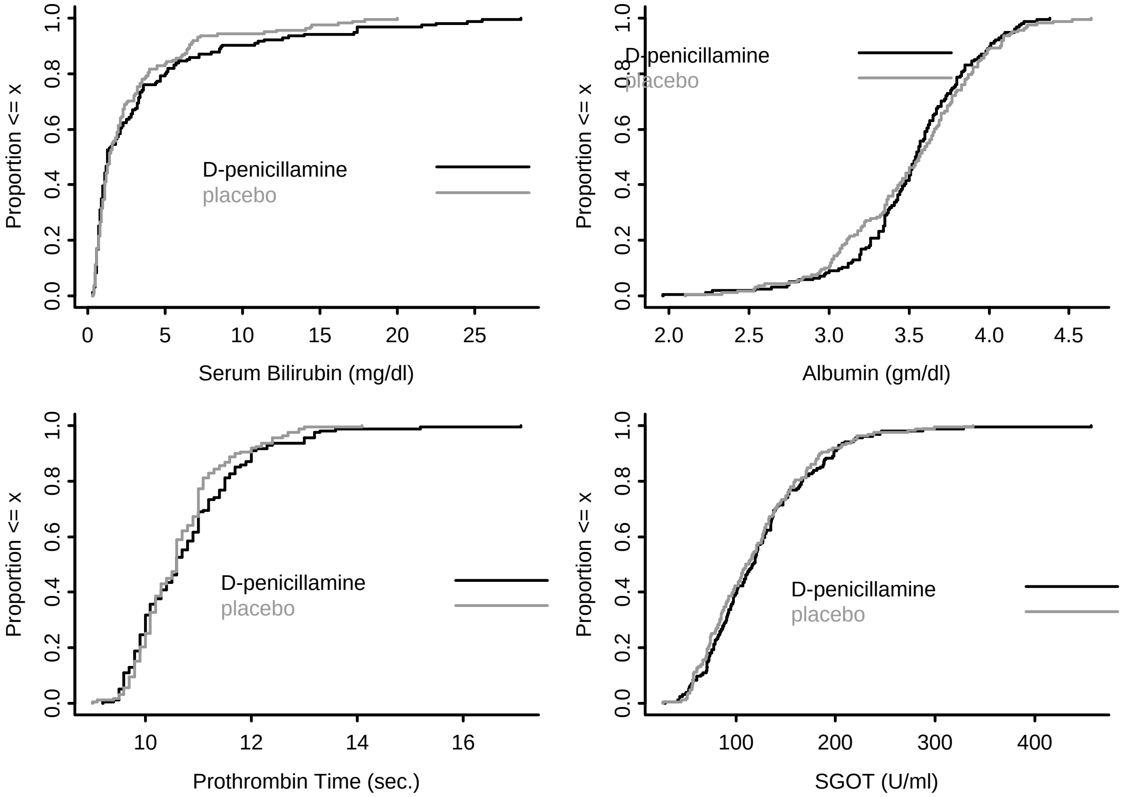


Figure 2.7: *Empirical cumulative distributions of baseline variables stratified by treatment in a randomized controlled trial.*

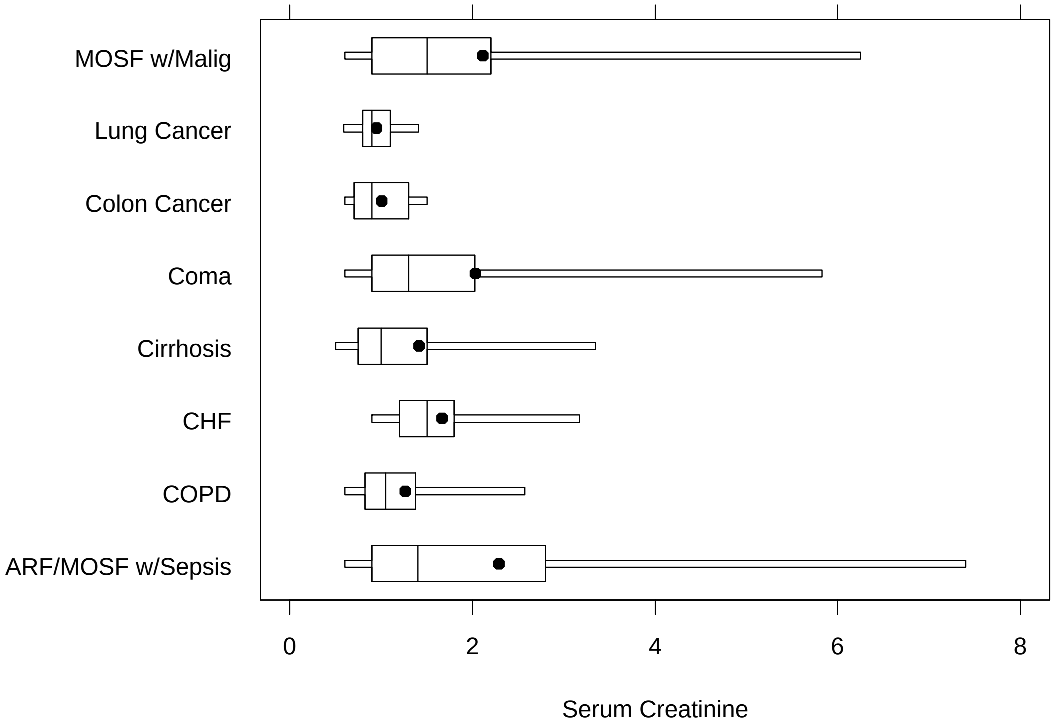


Figure 2.8: *Box plots showing the distribution of serum creatinine stratified by major diagnosis. Dot: mean; vertical line: median; large box: interquartile range. The 0.05 and 0.95 quantiles are also shown, which is not the way typical box plots are drawn but is perhaps more useful. Asymmetry of distributions can be seen by both disagreement between $Q_3 - Q_2$ and $Q_2 - Q_1$ and by disagreement between Q_2 and $\bar{x}$.*

Relationships

- When response variable is continuous and descriptor (stratification) variables are categorical, multi-panel dot charts, box plots, multiple cumulative distributions, etc., are useful.

- Two continuous variables: scatterplot (e.g., Rosner Figure 2.12, EMS Figure 3.9)

2.3.3 Graphs for Summarizing Results of Studies

- Dot charts with optional error bars (for confidence limits) can display any summary statistic (proportion, mean, median, mean difference, etc.)

- It is not well known that the confidence interval for a difference in two means cannot be derived from individual confidence limits.[c]
  Show individual confidence limits as well as actual CLs for the difference.

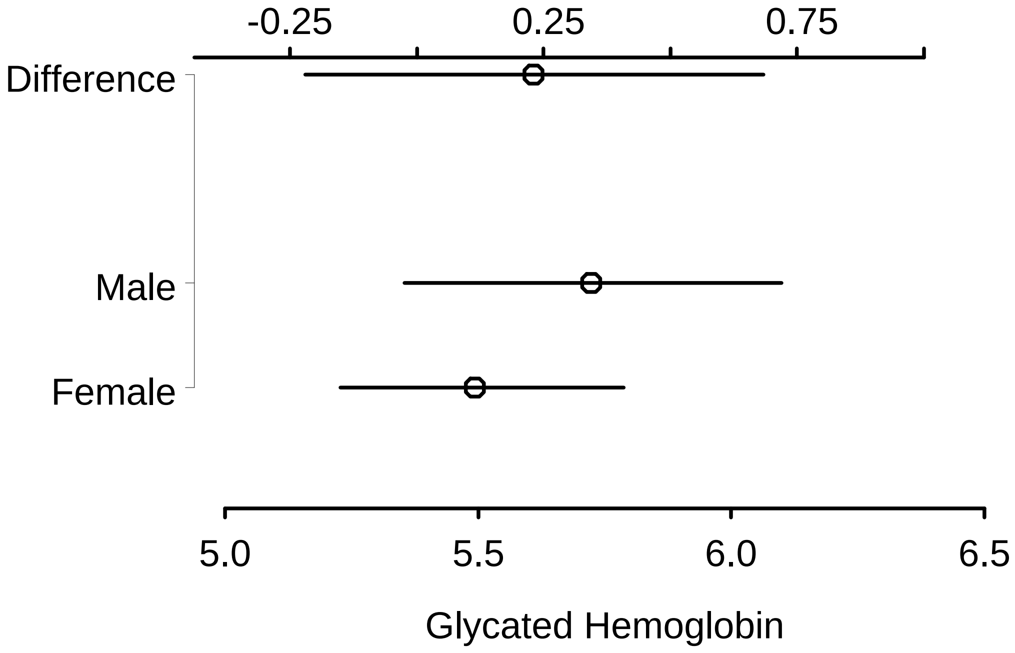


Figure 2.9: *Means and nonparametric bootstrap 0.95 confidence limits for glycated hemoglobin for males and females, and confidence limits for males - females. Lower and upper x-axis scales have same spacings but different centers. Confidence intervals for differences are generally wider than those for the individual constituent variables.*

- For showing relationship between two continuous variables, a trend line or regression model fit, with confidence bands

[c] In addition, it is not necessary for two confidence intervals to be separated for the difference in means to be significantly different from zero.

2.4 Tables

- Binary variables: don't need to show both proportions

- Make logical choices for independent and dependent variables.
  E.g., less useful to show proportion of males for patients who lived vs. those who died than to show proportion of deaths stratified by sex.

- Continuous variables
  - to summarize distributions of raw data: 3 quartiles
    recommended format: ${}_{35}$ **50** ${}_{67}$ or 35/50/67
  - summary statistics: mean or median and confidence limits (without assuming normality of data if possible)

- Show number of missing values

- Add denominators when feasible

Table 2.1: *Descriptive Statistics: Demographic and Clinical variables*

	N	
Age	27	${}_{28}$ 32 ${}_{52}$
C reactive protein	27	${}_{1.0}$ 1.8 ${}_{10.1}$
Fecal Calprotectin	26	${}_{128}$ 754 ${}_{2500}$
Gender	27	
Female		52% $\frac{14}{27}$
Location of colitis	27	
Left side		41% $\frac{11}{27}$
Middle		52% $\frac{14}{27}$
Right side		7% $\frac{2}{27}$

${}_{a}\, b\, {}_{c}$ represent the lower quartile a, the median b, and the upper quartile c for continuous variables.
N is the number of non–missing values.

2.5 Bar Plots with Error Bars

- “Dynamite” Plots

- Height of bar indicates mean, lines represent standard error

- High ink:information ratio

- Hide the raw data, assume symmetric confidence intervals

- Replace with
  - Dot plot (smaller sample sizes)

  - Box plot (larger sample size)

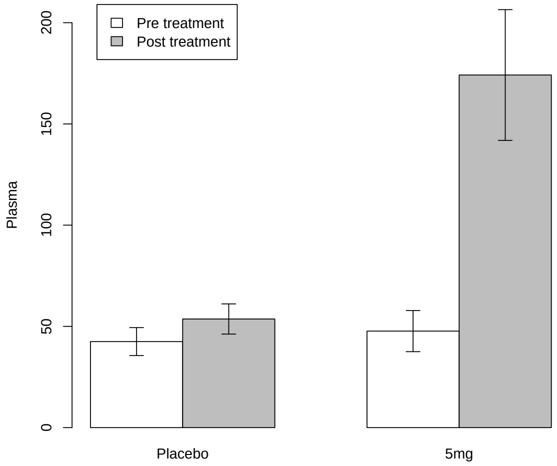


Figure 2.10: *Bar plot with error bars or “Dynamite” plot*

- Can Reproduce Placebo and 5mg in Figure 2.10 using two possible data structures
  - Symmetric plasma concentrations about the mean

  - “Responders” and “Non-responders” to folate

- Identical dynamite plots

- Differences shown in dot plots

- Other options

- Biostatistics TWiki
  - `http://biostat.mc.vanderbilt.edu/DynamitePlots`

2.5.1 Symmetric Observations

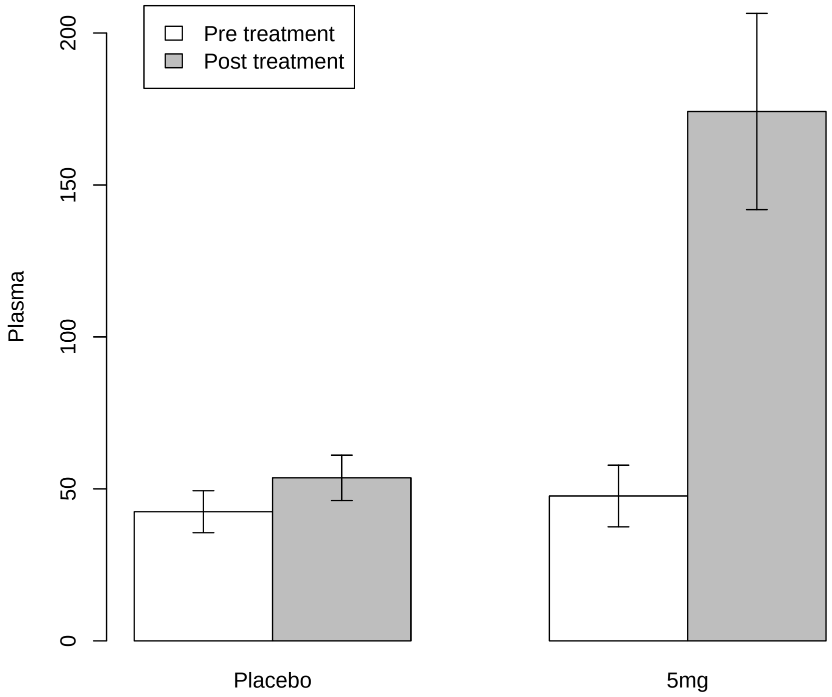


Figure 2.11: *Dynamite plot where underlying data is symmetric about the mean*

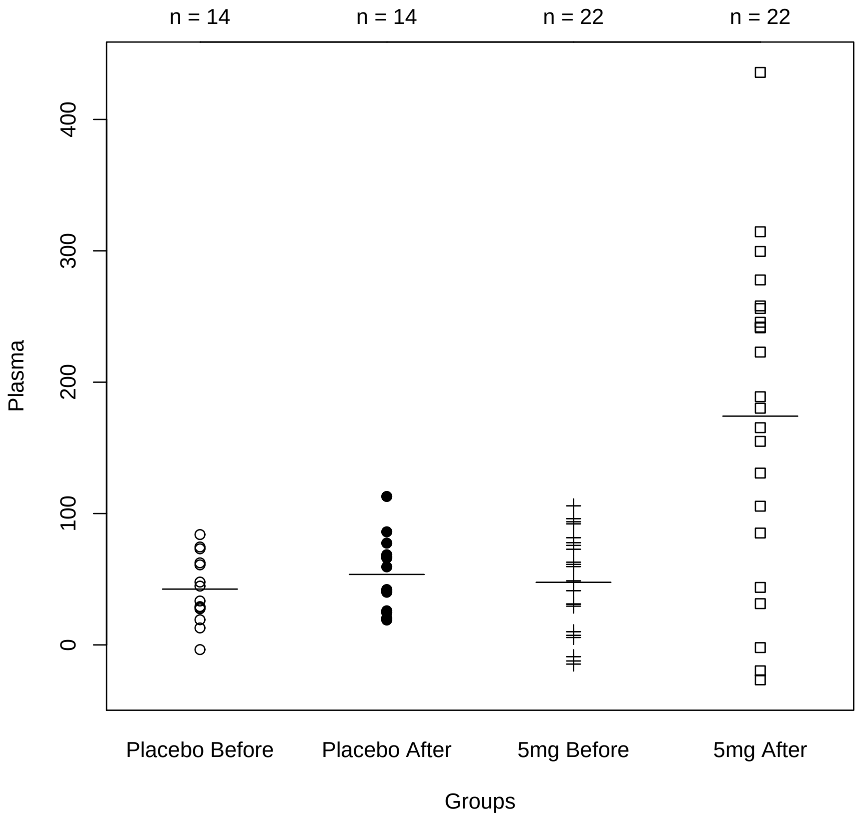


Figure 2.12: *Dot plot where underlying data is symmetric about the mean*

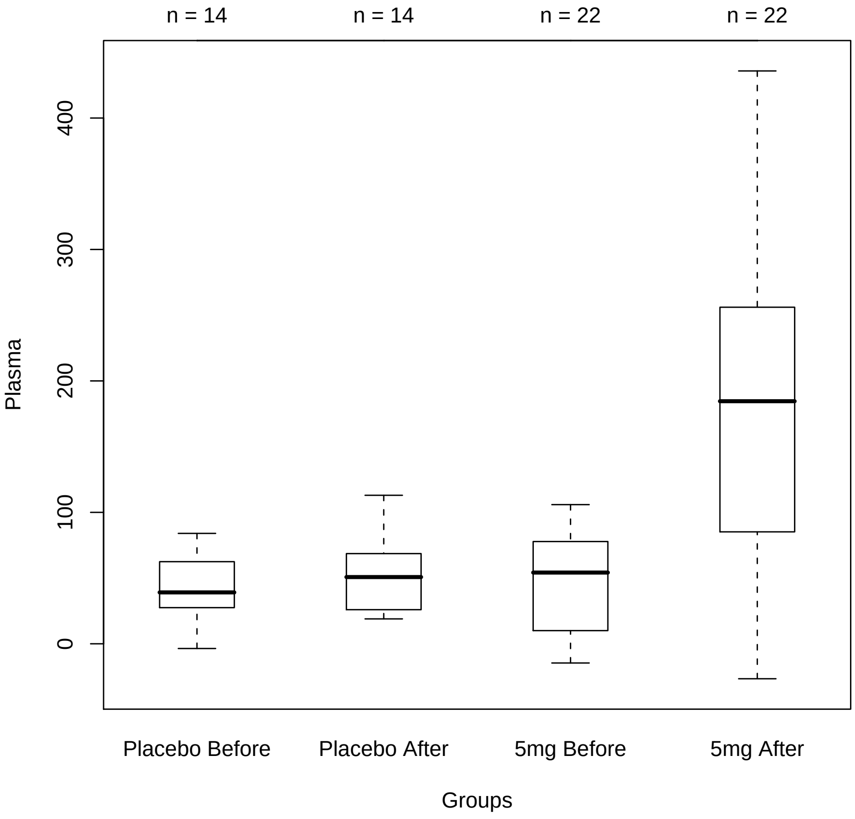


Figure 2.13: *Box plot where underlying data is symmetric about the mean*

2.5.2 Responder and Non-Responders

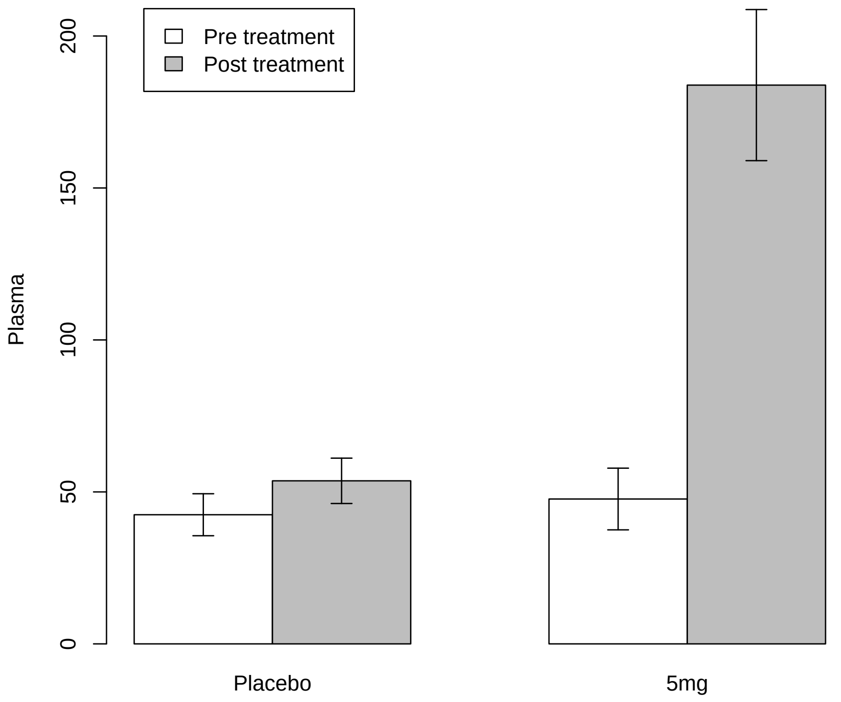


Figure 2.14: *Dynamite plot where underlying data contains responder and non-responders to folate*

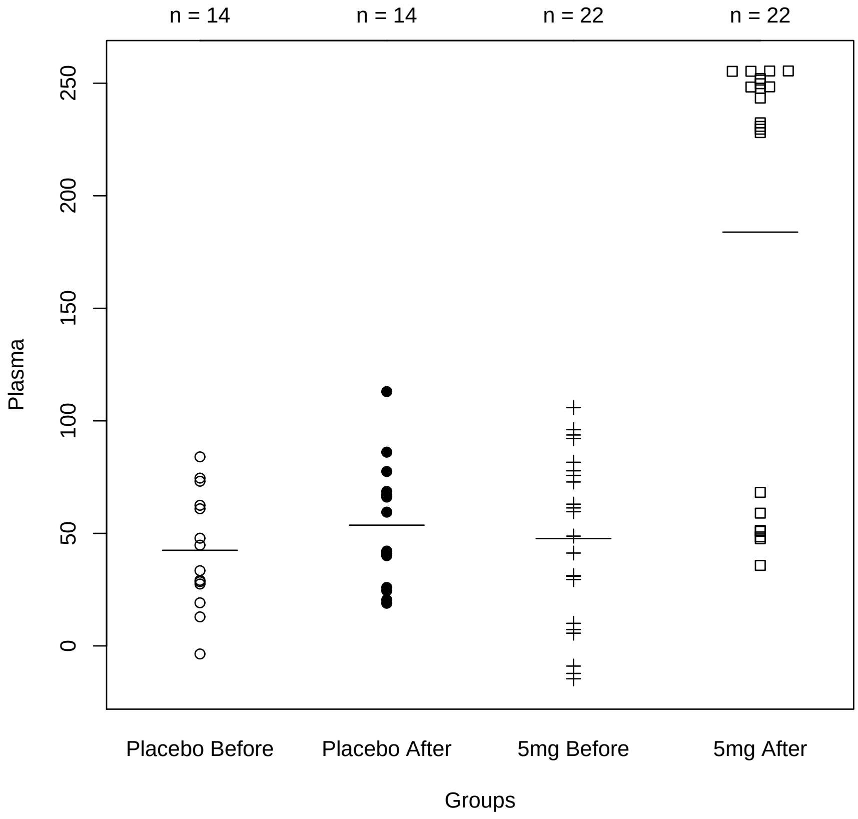


Figure 2.15: *Dot plot where underlying data contains responder and non-responders to folate*

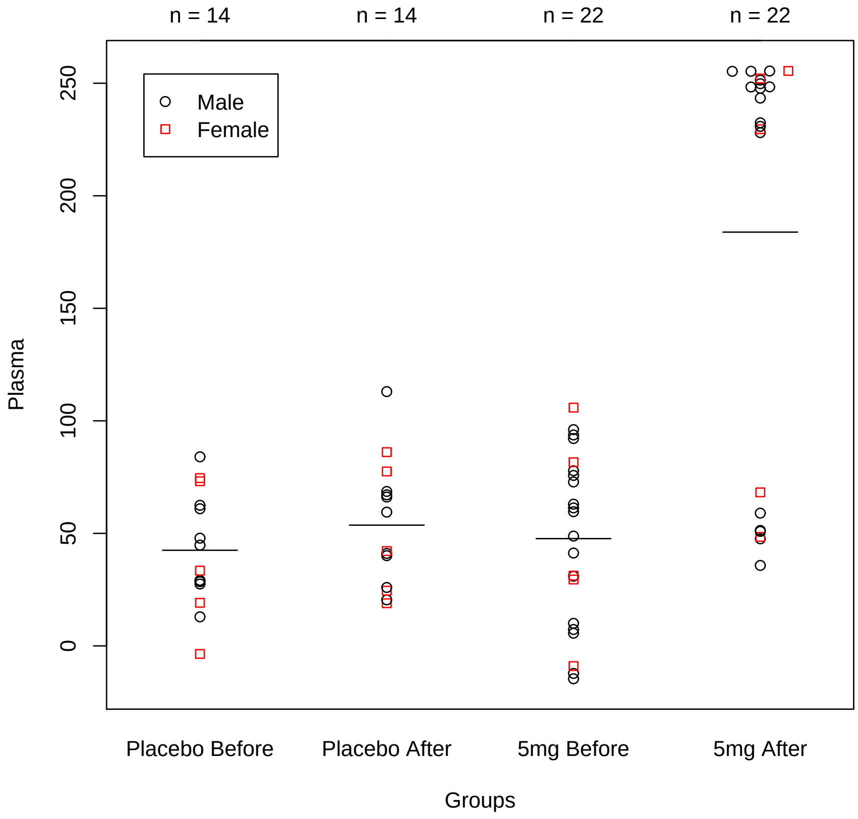


Figure 2.16: *Dot plot with different colors for gender*

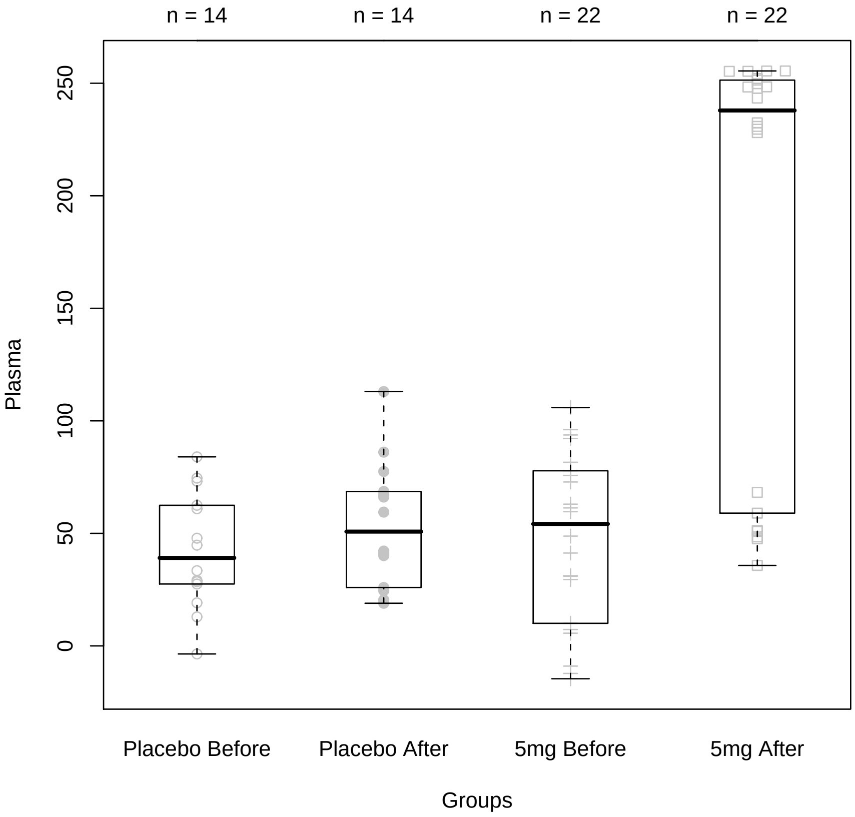


Figure 2.17: *Mixture of dot plot and box plot*

Chapter 3

Hypothesis Testing

3.1 Hypotheses

- Hypothesis: usually a statement to be judged of the form
  "population value = specified constant"
  - $\mu = 120$mmHg
  - $\mu_1 - \mu_2 = 0$mmHg
  - Correlation between wealth and religiosity = 0
- Null hypothesis is usually a hypothesis of no effect but can be $H_0 : \mu =$ constant or H_0 : Probability of heads $= \frac{1}{2}$;
  H_0 is often a straw man; something you hope to disprove
- Alternative hypothesis: H_1; e.g.: $H_1 : \mu \neq 120$mmHg
- One-sided hypothesis (tested by 1-tailed test): H_1 is an inequality in one direction ($H_1 : \mu > 120$mmHg)

- Two-sided hypothesis (2-tailed test, most common type): H_1 involves values far from the hypothesized value in either direction

3.2 Branches of Statistics

- Classical (frequentist or sampling statistics):
  - Emphasizes (overemphasizes?) hypothesis testing
  - Assumes H_0 is true
  - Conceives of data as one of many datasets that *might* have happened
  - See if data are consistent with H_0
  - Are data extreme or unlikely if H_0 is really true?
  - Proof by contradiction: if assuming H_0 is true leads to results that are “bizarre” or unlikely to have been observed, casts doubt on premise
  - Evidence summarized through a single statistic capturing a tendency of data, e.g., $\bar{x}$
  - Look at probability of getting a statistic as or more extreme than the calculated one (results as or more impressive than ours) if H_0 is true
  - If this statistic has a low probability of being observed to be this extreme we say that if H_0 is true we have acquired data that are very improbable, i.e., have witnessed a low probability event
  - Then evidence mounts against H_0 and we might reject it

  - A failure to reject *does not* imply that we have gathered evidence in favor of H_0 — many reasons for studies to not be impressive, including small sample size (n)

  - Ignores *clinical* significance

- Classical parametric statistics: assumes the data to arise from a certain distribution, often the normal (Gaussian distribution)

- Nonparametric statistics: does not assume a data distribution; generally looks at ranks rather than raw values

- Bayesian statistics:
  - Computes the probability that a clinically interesting statement is true, e.g. that the new drug lowers population mean SBP by at least 5mmHg, given what we observed in the data

  - More natural and direct approach but requires more work

  - Can formally incorporate knowledge from other studies as well as skepticism from a tough audience you are trying to convince to use a therapy

  - Starting to catch on (only been available for about 240 years) and more software becoming available

- We will deal with classical parametric and nonparametric statistical tests because of time

3.3 Errors in Hypothesis Testing

K7.5.A-7.5.B

- Can attempt to reject a formal hypothesis or just compute P-value

- Type I error: rejecting H_0 when it is true
  α is the probability of making this error (typically set at $\alpha = 0.05$—for weak reasons)

- Type II error: failing to reject H_0 when it is false
  probability of this is β

	True state of H_0	
Decision	H_0 true	H_0 false
Reject H_0	Type I error (α)	Correct
Do Not Reject H_0	Correct	Type II error (β)

- Power: $1 - \beta$: probability of (correctly) rejecting H_0 when it is false

A P-value is something that can be computed without speaking of errors. It is the probability of observing a statistic as or more extreme than the observed one if H_0 is true, i.e., if the population from which the sample was randomly chosen had the characteristics posited in the null hypothesis.

3.4 One Sample Test for Mean

3.4.1 Test

- Assuming continuous response from a normal distribution

- One sample tests for $\mu =$ constant are unusual except when data are paired, e.g., each patient has a pre– and post–treatment measurement and we are only interested in the mean of post - pre values

- t tests in general:

$$t = \frac{\text{estimate - hypothesized value}}{\text{standard deviation of numerator}}$$

- The standard deviation of a summary statistic is called its *standard error*, which is the $\surd$ of the variance of the statistic

- The one-sample t statistic for testing a single population mean against a constant μ_0 (H_0: $\mu = \mu_0$; often $\mu_0 = 0$) is
$$t = \frac{\bar{x} - \mu_0}{se}$$
where $se = \frac{s}{\sqrt{n}}$, is the standard error of the mean (SEM) and $\bar{x}$ is the sample mean

- When your data comes from a normal distribution and H_0 holds, the t ratio follows the t *distribution*

- With small sample size (n), the t ratio is unstable because the sample standard deviation (s) is not precise enough in estimating the population standard deviation (σ; we are assuming that σ is unknown)

- This causes the t distribution to have heavy tails for small n

- As $n \uparrow$ the t distribution becomes the normal distribution with mean zero and standard deviation one

- The parameter that defines the particular t distribution to use as a function of n is called the *degrees of freedom* or d.f.

- d.f. = n - number of means being estimated

- For one-sample problem d.f. = $n - 1$

- Symbol for distribution t_{n-1}

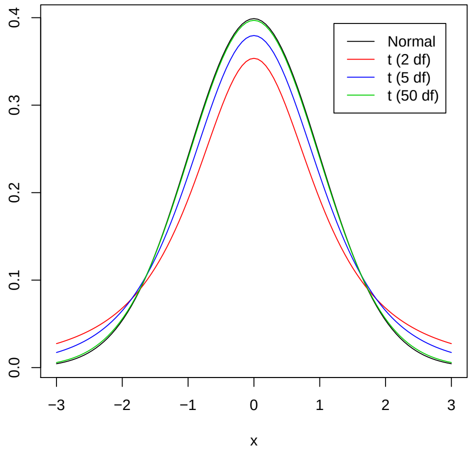


Figure 3.1: *Comparison of probability densities for t_2, t_5, t_{50}, and Normal distributions*

- Two-tailed P-value: probability of getting a value from the t_{n-1} distribution as big or bigger in absolute value than the absolute value of the observed t ratio

- Computer programs can compute the P-value given t and n.[a] See the course web site or go to `www.anu.edu.au/nceph/surfstat/surfstat-home/tables.html` for an interactive P and critical value calculator for common distributions.
  - don't say "$P <$ something" but instead $P =$ something

- In the old days tables were used to provide *critical values* of t, i.e., a value c of t such that Prob$[|t| > c] = \alpha$ for "nice" α such as 0.05, 0.01.

- Denote the critical value by $t_{n-1;1-\alpha/2}$ for a 2-tailed setup

- For large n (say $n \geq 500$) and $\alpha = 0.05$, this value approximates the value from the normal distribution, 1.96

- Example: We want to test if the mean tumor volume is 190 mm^3 in a popu-

[a] If the software has a function, say `tcdf` for the cumulative distribution function for the t distribution, the 2-tailed P-value would be obtained using a command like `2*(1-tcdf(abs(t),n-1))`.

lation with melanoma, $H_0 : \mu = 190$ versus $H_1 : \mu \neq 190$.

$$\begin{aligned} \bar{x} = 181.52, s = 40, n = 100, \mu_0 = 190 \\ t = \frac{181.52 - 190}{40/\sqrt{100}} = -2.12 \\ t_{99,.975} = 1.984 \rightarrow \text{reject at } \alpha = .05 \\ P = 0.037 \end{aligned}$$

3.4.2 Power and Sample Size

- Power ↑ when
  - allow larger type I error (α; tradeoff between type I and II errors)
  - true μ is far from μ_0
  - $\sigma \downarrow$
  - $n \uparrow$
- Power for 2-tailed test is a function of μ, μ_0 and σ only through $|\mu - \mu_0|/\sigma$
- Sample size to achieve $\alpha = 0.05$, power $= 0.9$ is approximately

$$n = 10.51 \left[\frac{\sigma}{\mu - \mu_0}\right]^2$$

- Some power calculators are at `statpages.org/#Power`
- Better: PS program by Dupont and Plummer `http://biostat.mc.vanderbilt.edu/PowerSampleSize`
- Example: The mean forced expiratory volume (FEV) in a population of asthmatics is 2.5 liters per second and the population standard deviation is assumed to be 1. Determine the number of subjects needed if a new drug is

expected to increase FEV to 3.0 liters per second ($\alpha = .05, \beta = 0.1$)

$$\mu = 2.5, \mu_0 = 3, \sigma = 1$$
$$n = 10.51 \left[\frac{1}{2.5 - 3}\right]^2 = 42.04$$

– Rounding up, we need 43 subjects to have 90% power (42 subjects would have less than 90% power)

3.4.3 Confidence Interval

A28-29

A 2-sided $1 - \alpha$ confidence interval for μ is

$$\bar{x} \pm t_{n-1,1-\alpha/2} \times se$$

The t constant is the $1 - \alpha/2$ level critical value from the t-distribution with $n - 1$ degrees of freedom. For large n it equals 1.96 when $\alpha = 0.05$.

A rough way to interpret this is that we are 0.95 confident that the unknown μ lies in the above interval. The exact way to say it is that if we were able to repeat the same experiment 1000 times and compute a fresh confidence interval for μ from each sample, we expect 950 of the samples to actually contain μ. Difficulties in providing exact interpretations of confidence intervals has driven many people to Bayesian statistics.

The 2-sided $1 - \alpha$ CL includes μ_0 if an only if a test of $H_0 : \mu = \mu_0$ is rejected at the α level in a 2-tailed test.

- If a 0.95 CL does not contain zero, we can reject $H_0 : \mu = 0$ at the $\alpha = 0.05$ significance level

$1 - \alpha$ is called the *confidence level* or *confidence coefficient*.

3.4.4 Sample Size for a Given Precision

A139-148

- May want to estimate μ to within a margin of error of $\pm\delta$ with 0.95 confidence

- “0.95 confident” that a confidence interval includes the true value of μ
- Width of confidence interval is 2δ

$$n = \left[\frac{t_{n-1,1-\alpha/2} se}{\delta}\right]^2$$

- If n is large enough and $\alpha = 0.05$, required $n = 3.84[\frac{se}{\delta}]^2$
- Example: if want to be able to nail down μ to within ± 1mmHg when the patient to patient standard deviation in blood pressure is 10mmHg, $n \sim 384$
- Advantages of planning for precision rather than power[b]
  - do not need to guess the true population value
  - many studies are powered to detect a miracle and nothing less; if a miracle doesn’t happen, the study provides **no** information
  - planning on the basis of precision will allow the resulting study to be interpreted if the P-value is large, because the confidence interval will not be so wide as to include both clinically significant improvement and clinically significant worsening

3.5 One Sample Method for a Probability

A:45-56

3.5.1 Test

- Estimate a population probability p with a sample probability $\hat{p}$
- Approximate 2-sided test of $H_0 : p = p_0$ obtained by computing a z statistic

[b] See Borenstein M: *J Clin Epi* 1994; 47:1277-1285.

- A z-test is a test assuming that the *test statistic* has a normal distribution; it is a t-test with infinite (∞) d.f.

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

- The z-test follows the same general form as the t-test

$$z = \frac{\text{estimate - hypothesized value}}{\text{standard deviation of numerator}}$$

- Example: $n = 10$ tosses of a coin, 8 heads; H_0: coin is fair ($p_0 = \frac{1}{2}$)

$$z = \frac{.8 - .5}{\sqrt{(\frac{1}{2})(\frac{1}{2})/10}} = 1.897$$

- $P\text{-value} = 2\times$ area under a normal curve to the right of $1.897 = 2 \times 0.0289 = 0.058$ (this is also the area under the normal curve to the right of 1.897 + the area to the left of -1.897)

- Approximate probability of getting 8 or more or 2 or fewer heads if the coin is fair is 0.058

- Need to use exact methods if p or n is small

3.5.2 Power and Sample Size

- Power $\uparrow$ as $n \uparrow$, p departs from p_0, or p_0 departs from $\frac{1}{2}$

- $n \downarrow$ as required power $\downarrow$ or p departs from p_0

3.5.3 Sample Size for Given Precision

- Approximate 0.95 CL: $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$

- Assuming p is between 0.3 and 0.8, it would not be far off to use the worst case standard error $\sqrt{1/(4n)}$ when planning

- n to achieve a margin of error δ in estimating p:

$$n = \frac{1}{4}\left[\frac{1.96}{\delta}\right]^2 = \frac{0.96}{\delta^2}$$

- Example: $\delta = .1 \rightarrow n = 96$ to achieve a margin of error of ± 0.1 with 0.95 confidence

3.6 Paired Data and One-Sample Tests

A31-32

- To investigate the relationship between smoking and bone mineral density, Rosner presented a paired analysis in which each person had a nearly perfect control which was his or her twin

- Data were normalized by dividing differences by the mean density in the twin pair (need to check if this normalization worked)

- Computed density in heavier smoking twin minus density in lighter smoking one

- Mean difference was -5% with se=2.0% on $n = 41$

- The t statistic we've been using works here, once within-pair differences are formed

- H_0 : mean difference between twins is zero ($\mu_0 = 0$)

$$t_{40} = \frac{\bar{x} - \mu_0}{se} = -2.5$$
$$P = 0.0166$$

3.7 Two Sample Test for Means

K5.5, A28-35

- Two groups of different patients (unpaired data)

- Much more common than one-sample tests

- As before we are dealing for now with parametric tests assuming the raw data arise from a normal distribution

- We assume that the two groups have the same spread or variability in the distributions of responses[c]

3.7.1 Test

- Test whether population 1 has the same mean as population 2

- Example: pop. 1=all patients with a certain disease if given the new drug, pop. 2=standard drug

- $H_0 : \mu_1 = \mu_2$ (this can be generalized to test $\mu_1 = \mu_2 + \delta$, i.e., $\mu_1 - \mu_2 = \delta$). The *quantity of interest* or *QOI* is $\mu_1 - \mu_2$

- 2 samples, of sizes n_1 and n_2 from two populations

- Two-sample (unpaired) t-test assuming normality and equal variances—recall that if we are testing against an H_0 of **no effect**, the form of the t test is

$$t = \frac{\text{point estimate of QOI}}{\text{se of numerator}}$$

- Point estimate QOI is $\bar{x}_1 - \bar{x}_2$

[c] Rosner covers the unequal variance case very well. As nonparametric tests have advantages for comparing two groups and are less sensitive to the equal spread assumption, we will not cover the unequal variance case here.

- Variance of the sum or difference of two independent means is the sum of the variance of the individual means

- This is $\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2[\frac{1}{n_1} + \frac{1}{n_2}]$

- Need to estimate the single σ^2 from the two samples

- We use a weighted average of the two sample variances:
$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- True standard error of the difference in sample means: $\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

- Estimate: $s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, so
$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- d.f. is the sum of the individual d.f., $n_1 + n_2 - 2$, where the -2 is from our having to estimate the center of two distributions

- If H_0 is true t has the $t_{n_1+n_2-2}$ distribution

- To get a 2-tailed P-value we compute the probability that a value from such a distribution is farther out in the tails of the distribution than the observed t value is (we ignore the sign of t for a 2-tailed test)

- Example: $n_1 = 8, n_2 = 21, s_1 = 15.34, s_2 = 18.23, \bar{x}_1 = 132.86, \bar{x}_2 = 127.44$
$$\begin{aligned} s^2 &= \frac{7(15.34)^2 + 20(18.23)^2}{7 + 20} = 307.18 \\ s &= \sqrt{307.18} = 17.527 \end{aligned}$$

$$\begin{aligned} se &= 17.527\sqrt{\frac{1}{8}+\frac{1}{21}} = 7.282 \\ t &= \frac{5.42}{7.282} = 0.74 \end{aligned}$$

on 27 d.f.

- $P = 0.466$ using the `Surfstat t-distribution calculator`
- Chance of getting a difference in means as larger or larger than 5.42 if the two populations really have the same means is 0.466
- $\rightarrow$ little evidence for concluding the population means are different

3.7.2 Power and Sample Size

- Power increases when
  - $\Delta = |\mu_1 - \mu_2| \uparrow$
  - $n_1 \uparrow$ or $n_2 \uparrow$
  - n_1 and n_2 are close
  - $\sigma \downarrow$
  - $\alpha \uparrow$
- Power depends on $n_1, n_2, \mu_1, \mu_2, \sigma$ approximately through

$$\frac{\Delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

- Note that when computing power using a program that asks for μ_1 and μ_2 you can just enter 0 for μ_1 and enter Δ for μ_2, as only the difference matters

- Often we estimate σ from pilot data, and to be honest we should make adjustments for having to estimate σ although we usually run out of gas at this point

- Easiest to use the power calculator at `statpages.org/#Power`

- Example:
  Get a pooled estimate of σ using $\sqrt{\frac{15.34^2+18.23^2}{2}} = 16.847$ when $\Delta = 5, n_1 = n_2 = 100, \alpha = 0.05$
  Program computed power of 0.550 compared to Rosner's 0.555 (the program on the Web page probably uses more accurate formulas that are difficult to use manually)

- Sample size depends on $k = \frac{n_2}{n_1}$, Δ, power, and α

- Sample size ↓ when
  - $\Delta \uparrow$
  - $k \rightarrow 1.0$
  - $\sigma \downarrow$
  - $\alpha \uparrow$
  - required power ↓

- An approximate formula for required sample sizes to achieve power $= 0.9$

with $\alpha = 0.05$ is

$$\begin{aligned} n_1 &= \frac{10.51\sigma^2(1+\frac{1}{k})}{\Delta^2} \\ n_2 &= \frac{10.51\sigma^2(1+k)}{\Delta^2} \end{aligned}$$

- Example using web page:
  Does not allow unequal n_1 and n_2; use power=0.8, $\alpha = 0.05, \mu_1 = 132.86, \mu_2 = 127.44, \sigma = 16.847$

- Result is 153 in each group (total=306) vs. Rosner's $n_1 = 108, n_2 = 216$, total=324. The price of having unequal sample sizes was 18 extra patients.

3.7.3 Confidence Interval

A28-35

$$\bar{x}_1 - \bar{x}_2 \pm t_{n_1+n_2-2,1-\alpha/2} \times s \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

is a $1-\alpha$ CL for $\mu_1 - \mu_2$, where s is the pooled estimate of σ, i.e., $s\sqrt{\ldots}$ is the estimate of the standard error of $\bar{x}_1 - \bar{x}_2$

3.7.4 Sample Size for a Given Precision

To design a study that will nail down the estimate of $\mu_1 - \mu_2$ to within $\pm\delta$ with $1-\alpha$ confidence when $n_1 = n_2 = n$, and when n is large enough so that the critical value $t_{2n-2,1-\alpha/2}$ may be approximated by the critical value from the normal distribution, say z ($z = 1.96$ when $\alpha = 0.05$):

$$n = 2\left[\frac{z\sigma}{\delta}\right]^2$$

When $\alpha = 0.05$, $n = 7.68[\frac{\sigma}{\delta}]^2$

3.7.5 Checking Assumptions of the t-test

- Box plot (one box for each of 2 groups): look for equal spread (IQR)
- Informally compare s_1 and s_2[d]
- Various plots for assessing normality of data from each group[e]

3.8 The Problem with Hypothesis Tests and P-values

3.8.1 Hypothesis Testing

- Existence of ESP is a hypothesis
- Assessing effects of drugs, procedures, devices involves estimation
- Many studies powered to detect huge effect
- If effect is not huge, no information from study

3.8.2 P-Values

A15-24

- Only provide evidence against a *null* hypothesis, **never** evidence for something
- Probability of a statistic as impressive as yours **if** H_0 true
- Not a probability of an effect or difference (same problem with sensitivity and specificity)

[d] Rosner 8.6 shows how to make formal comparisons, but beware that the variance ratio test depends on normality, and it may not have sufficient power to detect important differences in variances.

[e] There are formal tests of normality but in smaller samples these may have insufficient power to detect important nonnormality.

- **No** conclusion possible from large P-values

- Cannot conclude clinical relevance from small P

- Adjustment of P-values for multiple tests is controversial and there is insufficient consensus on how to choose an adjustment method

3.8.3 How Not to Present Results

- $P = 0.02$ — let's put this into clinical practice ignoring the drug's cost or clinical effectiveness

- $P = 0.4$ — this drug does not kill people

- $P = 0.2$ but there is a trend in favor of our blockbuster drug

- The observed difference was 6mmHg and we rejected H_0 so the true effect is 6mmHg.

- The proportion of patients having adverse events was 0.01 and 0.03; the study wasn't powered to detect adverse event differences so we present no statistical analysis

- The reduction in blood pressure was 6mmHg with 0.95 C.L. of [1mmHg, 11mmHg]; the drug is just as likely to only reduce blood pressure by 1mmHg as it is by 6mmHg.

- The serum pH for the 15 dogs was 7.3 ± 0.1 (mean $\pm$ SE)

3.8.4 How to Present Results

A15-24

- Estimates should be accompanied by confidence limits

- Confidence limits can be computed without regard to sample size or power

- A computed value from a sample is only an estimate of the population value, whether or not you reject H_0

- Best to think of an estimate from a study as a fuzz, not a point

- To present variability of subjects, use SD or IQR, **not** SE (SE is the precision of the *mean* of subjects)

3.9 Comprehensive Example: Two sample t-test

3.9.1 Study Description

- Compare the effects of two soporific drugs
  - Optical isomers of hyoscyamine hydrobromide

- Each subject receives a placebo and then is randomly assigned to receive Drug 1 or Drug 2

- Dependent variable: Number of hours of increased sleep over control

- Drug 1 given to n_1 subjects, Drug 2 given to n_2 different subjects

- Study question: Is Drug 1 or Drug 2 more effective at increasing sleep?
  - $H_0 : \mu_1 = \mu_2$

  - $H_1 : \mu_1 \neq \mu_2$

3.9.2 Power and Sample Size

- Pilot study or previous published research shows $\sigma = 1.9$ hours

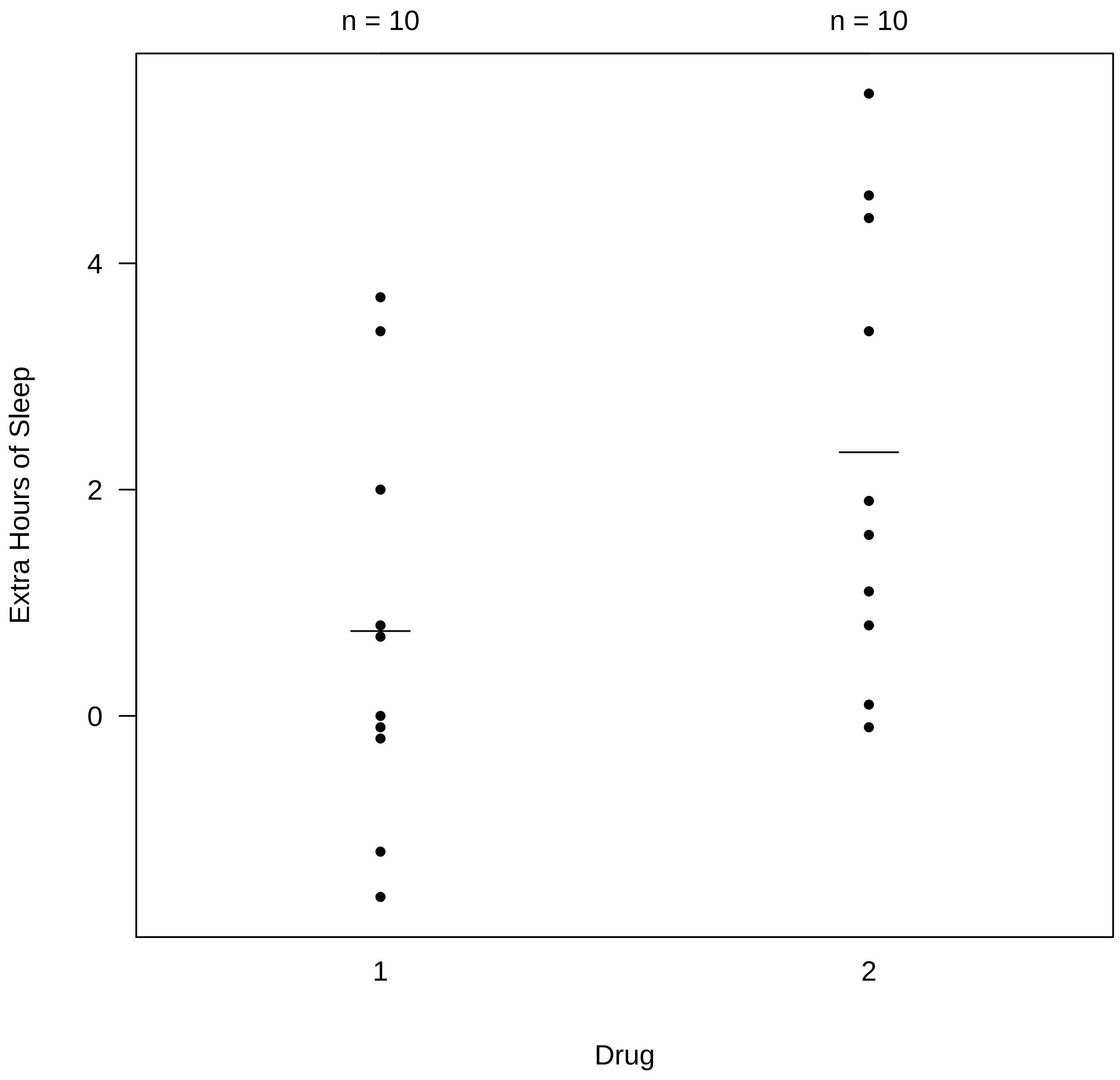
- Determine the number of subjects needed (in each group) for several value of effect size Δ ($\Delta = |\mu_1 - \mu_2|$) in order to have 90% power with $\alpha = 0.05$

Δ	1.0	1.5	2.0	2.5	3.0
n	77	35	20	14	10

- If Drug 1 (or 2) increases sleep by 3.0 hours more than Drug 2 (or 1), by enrolling 10 subjects in each group we will have 90% power to detect an association

3.9.3 Collected Data

Obs.	Drug 1	Drug 2
1	0.7	1.9
2	−1.6	0.8
3	−0.2	1.1
4	−1.2	0.1
5	−0.1	−0.1
6	3.4	4.4
7	3.7	5.5
8	0.8	1.6
9	0.0	4.6
10	2.0	3.4
Mean	0.75	2.33
SD	1.79	2.0


n = 10
n = 10
4
2
0
Extra Hours of Sleep
1
2
Drug

3.9.4 Statistical Test

- Stat program output

```
        Two Sample t-test

data:  extra by group
t = -1.8608, df = 18, p-value = 0.07919
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.3638740  0.2038740
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

- Interpretation
  - Compare Drug 2 to Drug 1. The output compares 1 to 2
  - Individuals who take Drug 2 sleep on average 1.58 hours longer (95% CI: [-0.20, 3.36]) than individuals who take Drug 1

3.10 Comprehensive Example: Paired t-test

3.10.1 Study Description

- Compare the effects of two soporific drugs.
- Each subject receives placebo, Drug 1, and Drug 2
- Dependent variable: Number of hours of increased sleep
- Drug 1 given to n subjects, Drug 2 given to same n subjects
- Study question: Is Drug 1 or Drug 2 more effective at increasing sleep?
  - $H_0 : \mu_d = 0 \text{ where } \mu_d = \mu_1 - \mu_2$
  - $H_1 : \mu_d \neq 0$

3.10.2 Power and Sample Size

- Pilot study or previous published research shows the standard deviation of the difference (σ_d) is 1.2 hours
- Determine the number of subjects needed for several value of effect size Δ ($\Delta = |\mu_1 - \mu_2|$)with 90% power, $\alpha = 0.05$

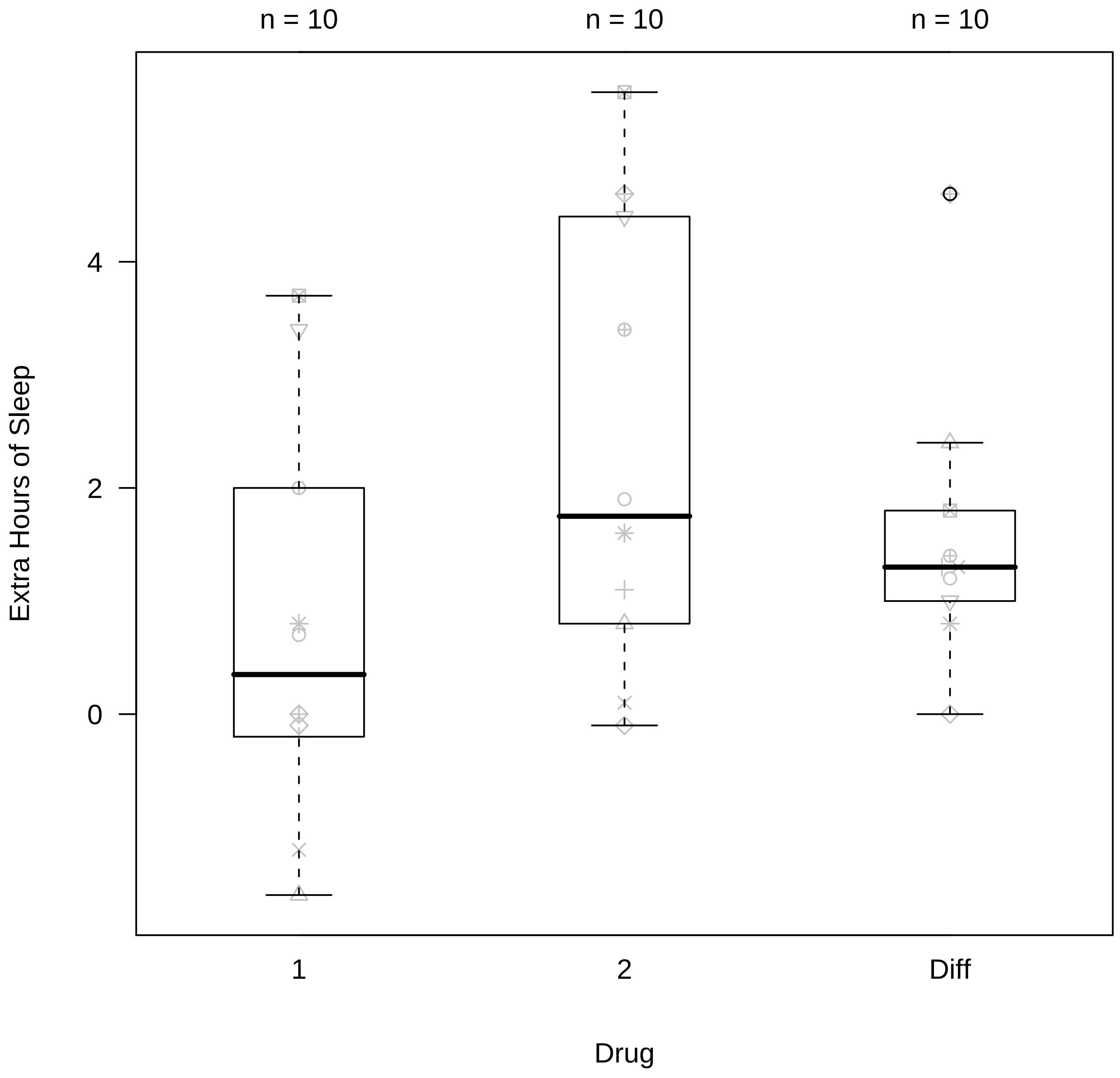
Δ	0.5	1	1.5	2
n	62	16	8	5

- If Drug 1 (or 2) increases sleep by 1.5 hours more than Drug 2 (or 1), by enrolling 8 subjects we will have 90% power to detect an association.

- More powerful than the two sample test (need 10 subjects in each group for $\Delta = 3.0$ hours)

3.10.3 Collected Data

Subject	Drug 1	Drug 2	Diff (2-1)
1	0.7	1.9	1.2
2	-1.6	0.8	2.4
3	-0.2	1.1	1.3
4	-1.2	0.1	1.3
5	-0.1	-0.1	0.0
6	3.4	4.4	1.0
7	3.7	5.5	1.8
8	0.8	1.6	0.8
9	0.0	4.6	4.6
10	2.0	3.4	1.4
Mean	0.75	2.33	1.58
SD	1.79	2.0	1.2


n = 10
n = 10
n = 10
Extra Hours of Sleep
4
2
0
1
2
Diff
Drug

3.10.4 Statistical Test

- Stat program output

```
         Paired t-test

data:  extra by group
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.4598858 -0.7001142
sample estimates:
mean of the differences
                  -1.58
```

- Interpretation
  - A person who takes Drug 2 sleeps on average 1.58 hours longer (95% CI: [0.70, 2.50]) than a person who takes Drug 1

- Note: Same point estimate (1.58 hours), but more precise estimate (tighter CI) than the 2-sample t-test

Chapter 4

Comparing Two Proportions

K5.2

4.1 Overview

- Compare dichotomous independent variable with a dichotomous outcome
  - Independent variables: Exposed/Not, Treatment/Control, Knockout/Wild Type, etc.
  - Outcome (dependent) variables: Diseased/Not or any Yes/No outcome
- Continuous outcomes often dichotomized for analysis (bad idea)
  - Consider t-tests (Chapter 3) or Non-parameteric methods (Chaper 5)

4.2 Normal-Theory Test

- Two independent samples

	Sample 1	Sample 2
Sample size	n_1	n_2
Population probability of event	p_1	p_2
Sample probability of event	$\hat{p}_1$	$\hat{p}_2$

- Null Hypothesis, $H_0 : p_1 = p_2 = p$

- Estimating the variance
  - Variance of $\hat{p}_i = p_i(1 - p_i)/n_i$ for $i = 1, 2$

  - Variance of $(\hat{p}_1 - \hat{p}_2)$ is the sum of the variances, which under H_0 is

    $$p(1 - p)[\frac{1}{n_1} + \frac{1}{n_2}]$$

  - We estimate this variance by plugging $\hat{p}$ into p, where

    $$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

    is the pooled estimate of the probability under $H_0 : p_1 = p_2 = p$

- Test statistic which has approximately a normal distribution under H_0 if $n_i\hat{p}_i$ are each large enough:

  $$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})[\frac{1}{n_1} + \frac{1}{n_2}]}}$$

- To test H_0 we see how likely it is to obtain a z value as far or farther out in the tails of the normal distribution than z is

- We don't recommend using the continuity correction

- Example:
  Test whether the population of women whose age at first birth ≤ 29 has the same probability of breast cancer as women whose age at first birth was ≥ 30. This dichotomization is highly arbitrary and we should really be testing for an association between age and cancer incidence, treating age as a continuous variable.

- Case-control study (independent and dependent variables interchanged); $p_1 =$ probability of age at first birth ≥ 30, etc.

	with Cancer	without Cancer
Total # of subjects	$3220 (n_1)$	$10245 (n_2)$
# age ≥ 30	683	1498
Sample probabilities	$0.212 (\hat{p}_1)$	$0.146 (\hat{p}_2)$
Pooled probability	$\frac{683+1498}{3220+10245} = 0.162$	

- Estimate the variance
  - $\text{variance}(\hat{p}_1 - \hat{p}_2) = \hat{p}(1-\hat{p}) \times \left[\frac{1}{n_1} + \frac{1}{n_2}\right] = 5.54 \times 10^{-5}$
  - $SE = \sqrt{\text{variance}} = 0.00744$
- Test statistic
  - $z = \frac{0.212-0.146}{0.00744} = 8.85$
- 2-tailed P-value is 0.0 using `survstat`; we report $P < 0.0001$
- We do not use a t-distribution because there is no σ to estimate (and hence no "denominator d.f." to subtract)

4.3 χ^2 Test

- If z has a normal distribution, z^2 has a χ^2 distribution with 1 d.f. (are testing a single difference against zero)
- The data we just tested can be shown as a 2×2 contingency table

	Cancer +	Cancer -	
Age ≤ 29	2537	8747	11284
Age ≥ 30	683	1498	2181
	3220	10245	13465

- In general, the χ^2 test statistic is given by

$$\sum_{ij} \frac{(\mathrm{Obs}_{ij} - \mathrm{Exp}_{ij})^2}{\mathrm{Exp}_{ij}}$$

- Obs_{ij} is the observed cell frequency for row i column j

- Exp_{ij} is the expected cell frequency for row i column j
  - Expected cell frequencies calculating assuming H_0 is true
  - $\mathrm{Exp}_{ij} = \frac{\text{row } i \text{ total} \times \text{column } j \text{ total}}{\text{grand total}}$
  - e.g. $\mathrm{Exp}_{11} = \frac{11284 \times 3220}{13465} = 2698.4$

- For 2×2 tables, if the observed cell frequencies are labeled $\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline \end{array}$ the χ^2 test statistic simplifies to

$$\frac{N[ad - bc]^2}{(a+c)(b+d)(a+b)(c+d)},$$

  where $N = a + b + c + d$. Here we get $\chi^2_1 = 78.37$

- 78.37 is z^2 from above!

- Don't need Yates' continuity correction Eq. 10.5

- Can get P-value from χ^2 distribution calculator (`surfstat`)

- Note that even though we are doing a 2-tailed test we use only the right tail of the χ^2_1 distribution; that's because we have squared the difference when computing the statistic, so the sign is lost.

- This is the ordinary Pearson χ^2 test

4.4 Fisher's Exact Test

- Is a misnomer in the sense that it computes probabilities exactly, with no normal approximation, but only after changing what is being tested to condition on the number of events and non-events

- As a result it is conservative

- The ordinary Pearson χ^2 works fine (even in most cases where an expected cell frequency < 5, contrary to popular belief)

- We don't use Yates' continuity correction because it was developed to make the normal approximation test yield P-values that are more similar to Fisher's test, i.e., to be more conservative

4.5 Sample Size and Power for Comparing Two Independent Samples

- Power $\uparrow$ as
  - $n_1, n_2 \uparrow$
  - $\frac{n_2}{n_1} \rightarrow 1.0$ (usually)
  - $\Delta = |p_1 - p_2| \uparrow$

  - $\alpha \uparrow$

- There are approximate formulas such as the recommended methods in Altman based on transforming $\hat{p}$ to make it have a variance that is almost independent of p A45-50

- Example:

  Using current therapy, 50% of the population is free of infection at 24 hours. Adding a new drug to the standard of care is expected to increase the percentage infection-free to 70%. If we randomly sample 100 subjects to receive standard care and 100 subjects to receive the new therapy, what is the probabilty that we will be able to detect a signficant different between the two therapies at the end of the study?

  $$p_1 = .5, p_2 = .7, n_1 = n_2 = 100$$

  results in a power of 0.83 when $\alpha = 0.05$

- When computing sample size to achieve a given power, the sample size $\downarrow$ when
  - power $\downarrow$
  - $\frac{n_2}{n_1} \rightarrow 1.0$
  - $\Delta \uparrow$
  - $\alpha \uparrow$

- Required sample size is a function of both p_1 and p_2

- Example:

How many subjects are needed to detect a 0.8 fold decrease in the probability of colorectal cancer if the baseline probability of cancer is 0.15%? Use a power of 0.8 and a type-I error rate of 0.05.

$$\begin{aligned} p_1 = 0.0015, p_2 = 0.8 \times p_1 = 0.0012, \alpha = 0.05, \beta = 0.2 \\ n_1 = n_2 = 234,945 \end{aligned}$$

according to `powercalc` (Rosner estimated 234,881)

4.6 Confidence Interval

An approximate $1-\alpha$ 2-sided CL is given by

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2} \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where $z_{1-\alpha/2}$ is the critical value from the normal distribution (1.96 when $\alpha = 0.05$).

The CL for the number of patients needed to be treated to save one event may simply be obtained by taking the reciprocal of the two confidence limits.[a]

4.7 Sample Size for a Given Precision

In the case $n_1 = n_2 = n, \alpha = 0.05$, the confidence interval has maximum width with p_1 and p_2 are near 0.5. As a worse case the margin of error is then $1.96/\sqrt{2n}$. The n required to achieve a margin of error of δ at the 0.95 confidence level is $1.92/\delta^2$. For example, to approximate the difference in the incidence probability of stroke between males and females to at worst ± 0.05 at the 0.95 level would require 768 patients in each group.

[a] If a negative risk reduction is included in the confidence interval, set the NNT to ∞ for that limit instead of quoting a negative NNT.

4.8 Relative Effect Measures

- We have been dealing with risk differences which are measures of absolute effect

- Measures of relative effect include risk ratios and odds ratios

- Risk ratios are easier to interpret but only are useful over a limited range of prognosis (i.e., a risk factor that doubles your risk of lung cancer cannot apply to a subject having a risk above 0.5 without the risk factor)

- Odds ratios can apply to any subject

- In large clinical trials treatment effects on lowering probability of an event are often constant on the odds ratio scale

- OR = Odds ratio = $\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$

- Testing H_0: OR=1 is equivalent to testing $H_0 : p_1 = p_2$

- There are formulas for computing confidence intervals for odds ratios

- Odds ratios are most variable when one or both of the probabilities are near 0 or 1

- We compute CLs for ORs by anti-logging CLs for the log OR

- In the case where $p_1 = p_2 = 0.05$ and $n_1 = n_2 = n$, the standard error of the log odds ratio is approximately $\sqrt{\frac{42.1}{n}}$

- The common sample size n needed to estimate the true OR to within a factor of 1.5 is 984 with ps in this range

4.9 Comprehensive example

4.9.1 Study Description

- Consider patients who will undergo coronary artery bypass graft surgery (CABG)
- Mortality risk associated with open heart surgery
- Study question: Do emergency cases have a surgical mortality that is different from that of non-emergency cases?
- Population probabilities
  - p_1: Probability of death in patients with emergency priority
  - p_2: Probability of death in patients with non-emergency priority
- Statistical hypotheses
  - $H_0 : p_1 = p_2$ (or $\mathrm{OR} = 1$)
  - $H_1 : p_1 \neq p_2$ (or $\mathrm{OR} \neq 1$)

4.9.2 Power and Sample Size

- Prior research shows that just over 10% of surgeries end in death
- Researchers want to be able to detect a 3 fold increase in risk

- For every 1 emergency priority, expect to see 10 non-emergency
- $p_1 = 0.3$, $p_2 = 0.1$, $\alpha = 0.05$, and $\mathrm{power} = 0.90$
- Calculate sample sizes using the PS software for these values and other combinations of p_1 and p_2

(p_1, p_2)	$\mathbf{(0.3, 0.1)}$	$(0.4, 0.2)$	$(0.03, 0.01)$	$(0.7, 0.9)$
n_1	**40**	56	589	40
n_2	**400**	560	5890	400

4.9.3 Collected Data

In-hospital mortality figures for emergency surgery and other surgery

	Discharge Status	
Surgical Priority	Dead	Alive
Emergency	6	19
Other	11	100

- $\hat{p}_1 = \frac{6}{25} = 0.24$
- $\hat{p}_2 = \frac{11}{111} = 0.10$

4.9.4 Statistical Test

- Stat program output

```
                 Discharge Status
 Priority Dead  Alive oddsratio    lower    upper  p.value
Emergency    6  19    1.000000       NA       NA   NA
Other       11  100   2.870813  0.946971 8.703085  0.05429

$measure
```

```
[1] "wald"

$conf.level
[1] 0.95

$pvalue
[1] "chi2"
```

- Interpretation
  - Compare odds of death in the emergency group $\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right)$ to odds of death in non-emergency group $\left(\frac{\hat{p}_2}{1-\hat{p}_2}\right)$
  - Emergency cases have 2.87 (95% CI: [0.95, 3.36]) fold increased odds times of death during surgery compared to non-emergency cases.

Fisher's Exact Test

Observed marginal totals from emergency surgery dataset

	Dead	Alive	
Emergency	a	b	25
Other	c	d	111
	17	119	136

- With fixed marginal totals, there are 18 possible tables ($a = 0, 1, \ldots 17$)
- Can calculated probability of each of these tables
  - p-value: Probability of observing data as extreme or more extreme than we collected in this experiment
- Exact test: p-value can be calculated "exactly" (not using the Chi-squared distribution to approximate the p-value)

- Stat program output

```
                  two-sided
Surgical Priority midp.exact fisher.exact chi.square
        Emergency         NA           NA         NA
        Other     0.07930086    0.0870594 0.05429257
```

- Fisher's test more conservative than Pearson's Chi-square Test (larger p-value)

Chapter 5

Nonparametric Statistical Tests

K5.4

5.1 When to use non-parametric methods

- Short answer: Good default when P-values are needed
- Nonparametric methods are those not requiring one to assume a certain distribution for the raw data
  - In contrast, parametric methods assume data come from some underlying distribution
  - t-tests assume the data come form a Normal distribution with parameters μ and σ^2 for the mean and variance, respectively
- Response variable ordinal or interval
- For ordinal responses nonparametric methods are preferred because they assume no spacing between categories
- No problem in using nonparametric tests on interval data

  - if normality holds, nonpar. test 0.95 efficient, i.e., has about the same power as the parametric test done on 0.95 of the observations
  - if normality does not hold, nonpar. tests can be arbitrarily more efficient and powerful than the corresponding parametric test
  - an elegant and non-arbitrary way to deal with extreme values or outliers
  - rank-based nonparametric tests give the analyst freedom from having to choose the correct transformation of the measurement (as long as the optimum transformation is monotonic)

- Example: Fecal calprotectin being evaluated as a possible biomarker of disease severity
  - Calprotectin has an upper detection limit
  - Median can be calculated (mean cannot)

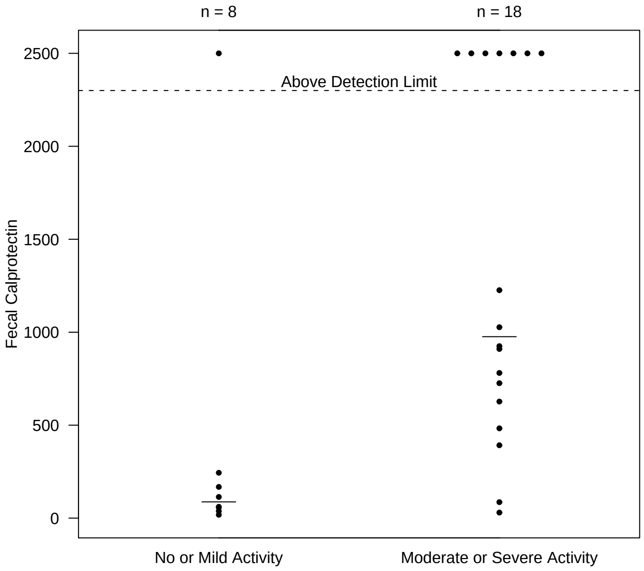


Figure 5.1: *Fecal calprotectin by endoscopy severity rating*

- If all you want is a P-value nonpar. tests are preferred

- A drawback is that nonpar. tests do not correspond to usual confidence limits for effects
  - E.g., a CL for the difference in 2 means may include zero whereas the Wilcoxon test yields $P = 0.01$
  - Point estimate that exactly corresponds to the Wilcoxon two-sample test is the Hodges-Lehman estimate of the location difference
    - $*$ median of all possible differences between a measurement from group 1 and a measurement from group 2

- Nonparametric tests are often obtained by replacing the data with ranks across subjects and then doing the parametric test

- Many nonpar. tests give the same P-value regardless of how the data are transformed; a careful choice of transformation (e.g., log) must sometimes be used in the context of parametric tests

- P-values computed using e.g. the t distribution are quite accurate for nonparametric tests

- In case of ties, midranks are used, e.g., if the raw data were 105 120 120 121 the ranks would be 1 2.5 2.5 4

Parametric Test	Nonparametric Counterpart
1-sample t	Wilcoxon signed-rank
2-sample t	Wilcoxon 2-sample rank-sum
k-sample ANOVA	Kruskal-Wallis
Pearson r	Spearman ρ

5.2 One Sample Test: Wilcoxon Signed-Rank

K5.10.E

- Almost always used on paired data where the column of values represents differences (e.g., post-pre) or log ratios

- The *sign test* is the simplest test for the median difference being zero in the population
  - it just counts the number of positive differences after tossing out zero differences
  - tests H_0 :Prob$[x > 0] = \frac{1}{2}$, i.e., that it is equally likely in the population to have a value below zero as it is to have a value above zero
  - this is the same as testing that the population median difference is zero
  - as it ignores magnitudes completely, the test is inefficient

- In the Wilcoxon signed rank one-sample test, ranks of absolute differences are given the sign of the original difference

- Magnitudes of raw data matter more here than with the Wilcoxon 2-sample test

- Observations with zero differences are ignored

- Example: A crossover study in which the treatment order is randomized Data arranged so that treatment A is in the first column, no matter which order treatment A was given

A	B	B-A	Rank $\lvert \mathrm{B}-\mathrm{A} \rvert$	Signed Rank
5	6	1	1.5	1.5
6	5	-1	1.5	-1.5
4	9	5	4.0	4.0
7	9	2	3.0	3.0

- A good approximation to an exact P-value may be obtained by computing

  $$z = \frac{\sum SR_i}{\sqrt{\sum SR_i^2}},$$

  where the signed rank for observation i is SR_i. This formula already takes ties into account without using Rosner's messy Eq. 9.5. We look up $|z|$ against the normal distribution. Here $z = \frac{7}{\sqrt{29.5}} = 1.29$ and by `surfstat` the 2-tailed P-value is 0.197

- If all differences are positive or all are negative, the exact 2-tailed P-value is $\frac{1}{2^{n-1}}$
  - implies that n must exceed 5 for any possibility of significance at the $\alpha = 0.05$ level for a 2-tailed test

5.2.1 One sample/Paired Test Example

- Sleep Dataset
  - Compare the effects of two soporific drugs.
  - Each subject receives placebo, Drug 1, and Drug 2
  - Study question: Is Drug 1 or Drug 2 more effective at increasing sleep?
  - Dependent variable: Difference in hours of sleep comparing Drug 2 to Drug 1

- H_0 : For any given subject, the difference in hours of sleep is equally likely to be positive or negative

Subject	Drug 1	Drug 2	Diff (2-1)	Sign	Rank
1	1.9	0.7	−1.2	-	3
2	−1.6	0.8	2.4	+	8
3	−0.2	1.1	1.3	+	4.5
4	−1.2	0.1	1.3	+	4.5
5	−0.1	−0.1	0.0	NA	NA
6	3.4	4.4	1.0	+	2
7	3.7	5.5	1.8	+	7
8	0.8	1.6	0.8	+	1
9	0.0	4.6	4.6	+	9
10	2.0	3.4	1.4	+	6

Table 5.1: *Hours of extra sleep on drugs 1 and 2, differences, signs and ranks of sleep study data*

- Wilcoxon signed rank test statistical program output

```
        Wilcoxon signed rank test

data:  sleep.data
V = 42, p-value = 0.02077
alternative hypothesis: true location is not equal to 0
```

  – Interpretation: Reject H_0, Drug 2 increases sleep by more hours than Drug 1 ($p = 0.02$)

- Could also perform sign test on sleep data

  – If drugs are equally effective, should have same number of '+' and '-'

  – Observed data: 1 '-', 8 '+', throw out 1 'no change'

  – Sign test (2-sided) P-value: Probability of observing 0 or 1 '-' OR 0 or 1 '+'

  – $p = 0.04$, so reject H_0 at $\alpha = 0.05$

- The signed rank test assumes that the distribution of differences is symmetric

- When the distribution is symmetric, the signed rank test tests whether the median difference is zero

- In general it tests that, for two randomly chosen observations i and j with values (differences) x_i and x_j, that the probability that $x_i + x_j > 0$ is $\frac{1}{2}$

- The estimator that corresponds exactly to the test in all situations is the pseudomedian, the median of all possible pairwise averages of x_i and x_j, so one could say that the signed rank test tests H_0: pseudomedian=0

- The value $\frac{\overline{SR}}{n+1} - \frac{1}{2}$ estimates the probability that two randomly chosen observations have a positive sum, where $\overline{SR}$ is the mean of the column of signed ranks

- To test $H_0 : \eta = \eta_0$, where η is the population median (not a difference) and η_0 is some constant, we create the n values $x_i - \eta_0$ and feed those to the signed rank test, assuming the distribution is symmetric

- When all nonzero values are of the same sign, the test reduces to the *sign test* and the 2-tailed P-value is $(\frac{1}{2})^{n-1}$ where n is the number of nonzero values

5.3 Two Sample Test: Wilcoxon–Mann–Whitney

K5.5.B

- The Wilcoxon–Mann–Whitney (WMW) 2-sample rank sum test is for testing for equality of central tendency of two distributions (for unpaired data)

- Ranking is done by combining the two samples and ignoring which sample each observation came from

- Example:

Females	120	118	121	119
Males	124	120	133	
Ranks for Females	3.5	1	5	2
Ranks for Males	6	3.5	7	

- Doing a 2-sample t-test using these ranks as if they were raw data and computing the P-value against 4+3-2=5 d.f. will work quite well

- Some statistical packages compute P-values exactly (especially if there are no ties)

- Loosely speaking the WMW test tests whether the population medians of the two groups are the same

- More accurately and more generally, it tests whether observations in one population tend to be larger than observations in the other

- Letting x_1 and x_2 respectively be randomly chosen observations from populations one and two, WMW tests $H_0 : C = \frac{1}{2}$, where $C =$Prob$[x_1 > x_2]$

- The C index (*concordance probability*) may be estimated by computing

  $$C = \frac{\bar{R} - \frac{n_1+1}{2}}{n_2},$$

  where $\bar{R}$ is the mean of the ranks in group 1;
  For the above data $\bar{R} = 2.875$ and $C = \frac{2.875-2.5}{3} = 0.125$, so we estimate that the probability is 0.125 that a randomly chosen female has a value greater than a randomly chosen male.

- In diagnostic studies where x is the continuous result of a medical test and the grouping variable is diseased vs. nondiseased, C is the area under the receiver operating characteristic (ROC) curve

- Test still has the “probability of ordering” interpretation when the variances of the two samples are markedly different, but it no longer tests anything like the difference in population medians

5.3.1 Two sample WMW example

- Fecal calprotectin being evaluated as a possible biomarker of disease severity
- Calprotectin measured in 26 subjects, 8 observed to have no/mild activity by endoscopy
- Calprotectin has upper detection limit at 2500 units
  - A type of missing data, but need to keep in analysis
- Study question: Are calprotectin levels different in subjects with no or mild activity compared to subjects with moderate or severe activity?
- Statement of the null hypothesis
  - H_0 : Populations with no/mild activity have the same distribution of calprotectin as populations with moderate/severe activity
  - $H_0 : C = \frac{1}{2}$
- Stat program output

```
        Wilcoxon rank sum test

data:  calpro by endo2
W = 23.5, p-value = 0.006257
alternative hypothesis: true location shift is not equal to 0
```

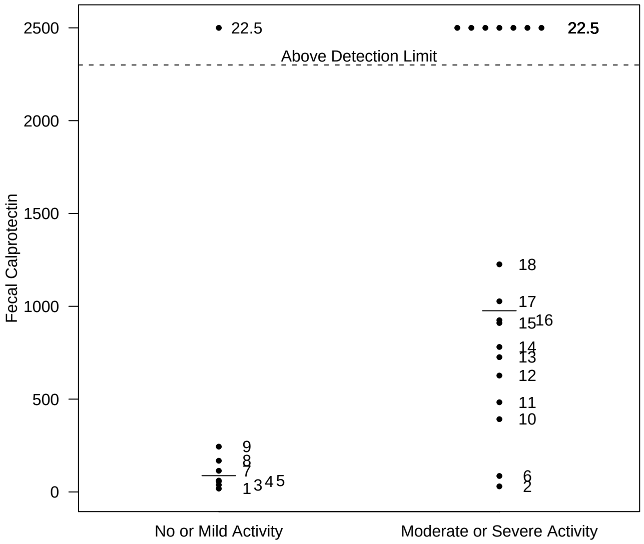


Figure 5.2: *Fecal calprotectin by endoscopy severity rating. Numbers indicate ranks of calprotectin levels, ignoring group*

- Test statistic W equals the sum of the ranks in the no/mild group minus $n_1 * (n_1 + 1)/2$, where n_1 is the number of subjects in then no/mild sample

- $W = 59.5 - \frac{8*9}{2} = 23.5$

- A common (but loose) interpretation: People with moderate/severe activity have higher *median* fecal calprotectin levels than people with no/mild activity ($p = 0.006$).

5.4 Confidence Intervals

A36-43

- Confidence intervals for the median (one sample)
  - Table 18.4 (Altman) gives the ranks of the observations to be used to give approximate confidence intervals for the median

  - e.g., if $n = 12$, the 3^{rd} and 10^{th} largest values give a 96.1% confidence interval

  - For larger sample sizes, the lower ranked value (r) and upper ranked value (s) to select for an approximate 95% confidence interval for the population median is

$$r = \frac{n}{2} - 1.96 * \frac{\sqrt{n}}{2} \quad \text{and} \quad s = 1 + \frac{n}{2} + 1.96 * \frac{\sqrt{n}}{2}$$

  - e.g., if $n = 100$ then $r = 40.2$ and $s = 60.8$, so we would pick the 40^{th} and 61^{st} largest values from the sample to specify a 95% confidence interval for the population median

- Confidence intervals for the difference in two medians (two samples)

  - Assume data come from distributions with same shape and differ only in location

  - Considers all possibly differences between sample 1 and sample 2

	Female			
Male	120	118	121	119
124	4	6	3	5
120	0	2	-1	1
133	13	15	12	14

  - An estimate of the median difference (males - females) is the median of these 12 differences, with the 3^{rd} and 10^{th} largest values giving an (approximate) 95% CI

  - Median estimate = 4.5, 95% CI = [1, 13]

  - Specific formulas found in Altman, pages 40-41

- Bootstrap A159-163
  - General method, not just for medians
  - Non-parametric, does not assume symmetry
  - Iterative method that repeatedly samples from the original data
  - Algorithm for creating a 95% CI for the difference in two medians
    1. Sample *with replacement* from sample 1 and sample 2
    2. Calculate the difference in medians, save result
    3. Repeat Steps 1 and 2 1000 times
  - A (naive) 95% CI is given by the 25^{th} and 975^{th} largest values of your 1000 median differences
  - For the male/female data, median estimate = 4.5, 95% CI = [-0.5, 14.5], which agrees with the conclusion from a WMW rank sum test ($p = 0.11$).

5.5 Strategy

- Don't assess normality of data
- Use nonparametric test in any case, to get P-values
- Use nonparametric confidence intervals for means and medians[a] which will be more in conformance to what the nonpar. test is testing

[a] A good nonparametric confidence for a population mean that does not even assume a symmetric distribution can be obtained from the bootstrap simulation procedure.

Chapter 6

Correlation

6.1 Overview

Outcome	Predictor	Normality?	Analysis Method
Interval	Binary	Yes	T-tests **or linear regression**
Interval	Binary	No	Wilcoxon 1- and 2-sample tests
Categorical	Categorical	NA	Pearson's Chi-square test
Interval	Interval	Yes	**Correlation or linear regression**
Interval	Interval	No	**Spearman's rank correlation**

- Examine association between continuous/interval outcome (y) and continous/interval predictor (x)

- Scatterplot of y versus x

6.2 Pearson's correlation coefficient

K:5.7.A

- $r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}$

- Range: $-1 \leq r \leq 1$

- Correlation coefficient is a unitless index of strength of association between two variables (+ = positive association, - = negative, 0 = no association)

- Measures the linear relationship between X and Y

- Can test for significant association by testing whether the population correlation is zero
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$
which is identical to the t-test used to test whether the population r is zero; d.f.=$n-2$.

- Use probability calculator for t distribution to get P-value (2-tailed if interested in association in either direction)

- 1-tailed test for a positive correlation between X and Y tests $H_0:$ when $X \uparrow$ does $Y \uparrow$ in the population?

- Confidence intervals for population r calculated using Fisher's Z transformation A:89-91
$$Z = \frac{1}{2}\log_{\mathrm{e}}\left(\frac{1+r}{1-r}\right)$$
  - For large n, Z follows a Normal distribution with standard error $\frac{1}{\sqrt{n-3}}$
  - To calculate a confidence interval for r, first find the confidence interval for Z then transform back to the r scale
$$\begin{aligned} Z &= \frac{1}{2}\log_{\mathrm{e}}\left(\frac{1+r}{1-r}\right) \\ 2*Z &= \log_{\mathrm{e}}\left(\frac{1+r}{1-r}\right) \end{aligned}$$

$$
\begin{aligned}
\exp(2*Z) &= \left(\frac{1+r}{1-r}\right) \\
\exp(2*Z)*(1-r) &= 1+r \\
\exp(2*Z) - r*\exp(2*Z) &= 1+r \\
\exp(2*Z) - 1 &= r*\exp(2*Z) + r \\
\exp(2*Z) - 1 &= r\,(\exp(2*Z) + 1) \\
\frac{\exp(2*Z)-1}{\exp(2*Z)+1} &= r
\end{aligned}
$$

- Example (Altman 89-90): Pearson's r for a study investigating the association of basal metabolic rate with total energy expenditure was calculated to be 0.7283 in a study of 13 women. Derive a 95% confidence interval for r.

  $$Z = \frac{1}{2}\log_e\left(\frac{1+0.7283}{1-0.7283}\right) = 0.9251$$

  The lower limit of a 95% CI for Z is given by

  $$0.9251 - 1.96 * \frac{1}{13-3} = 0.3053$$

  and the upper limit is

  $$0.9251 + 1.96 * \frac{1}{13-3} = 1.545$$

  A 95% CI for the population correlation coefficient is given by transforming these limits from the Z scale back to the r scale

  $$\frac{\exp(2*0.3053)-1}{\exp(2*0.3053)+1} \quad \text{to} \quad \frac{\exp(2*1.545)-1}{\exp(2*1.545)+1}$$

  Which gives a 95% CI from 0.30 to 0.91 for the population correlation

6.3 Spearman's Rank Correlation

K:5.7.B

- Pearson's r assumes linear relationship between X and Y

- Spearman's ρ (sometimes labeled r_s) assumes monotonic relationship between X and Y
  - when $X \uparrow$, Y always $\uparrow$ or stays flat, or Y always $\downarrow$ or stays flat
  - does not assume linearity
- $\rho = r$ once replace column of Xs by their ranks and column of Ys by ranks
- To test $H_0 : \rho = 0$ without assuming linearity or normality, being damaged by outliers, or sacrificing much power (even if data are normal), use a t statistic:

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$$

  which is identical to the t-test used to test whether the population r is zero; d.f.=$n-2$.
- Use probability calculator for t distribution to get P-value (2-tailed if interested in association in either direction)
- 1-tailed test for a positive correlation between X and Y tests H_0 : when $X \uparrow$ does $Y \uparrow$ in the population?

6.4 Correlation Examples

- Correlation difficult to judge by eye
- Example plots on following pages

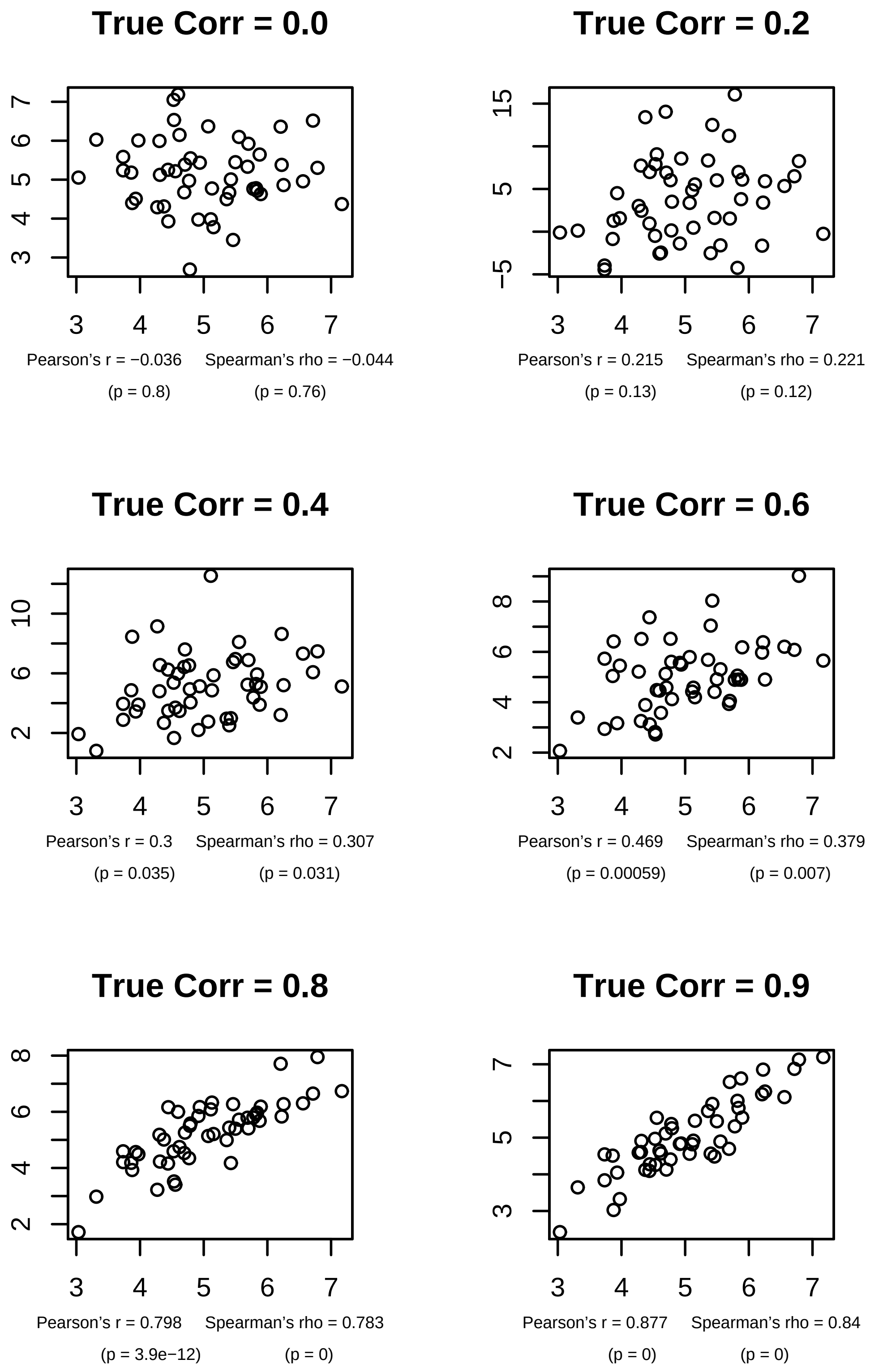


Figure 6.1: *X and Y are drawn from bivariate Normal populations with correlations ranging from 0.0 to 0.9. Pearson and Spearman sample correlations are shown for samples of size 50.*

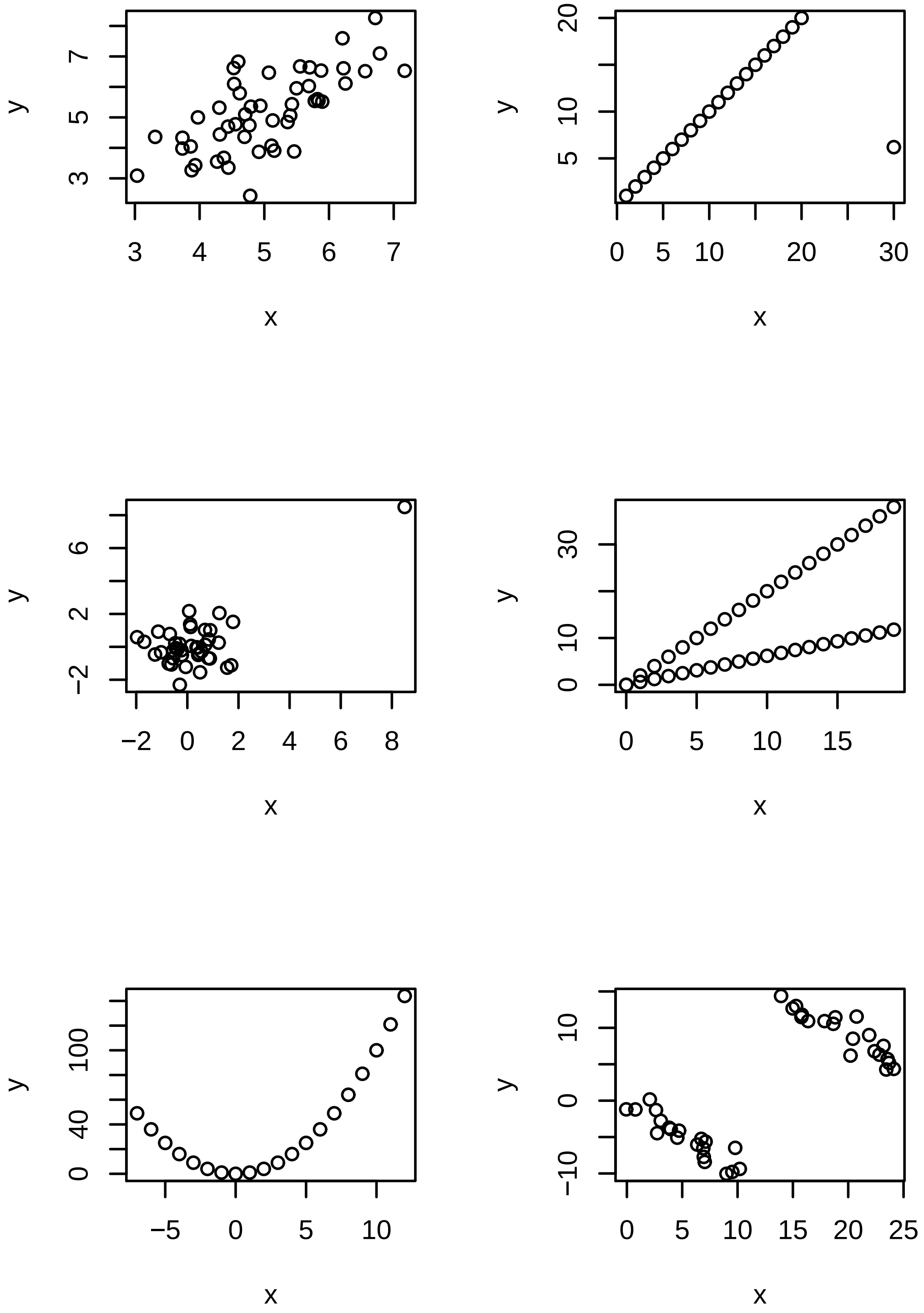


Figure 6.2: *Different observed datasets that have the same correlation. All six plots have a sample Pearson's correlation of* 0.7*.*

6.5 Correlation and Agreement

- Compare two methods of measuring the same underlying value
  - Lung function measured using a spirometer (expensive, accurate) or peak flow meter (cheap, less accurate)
  - Two devices (Restech and Sandhill) used to mesured acidity (pH) in the esophagus
- Typical (incorrect) approach begins with scatterplot of Restech versus Sandhill with a 1:1 line indicating perfect agreement

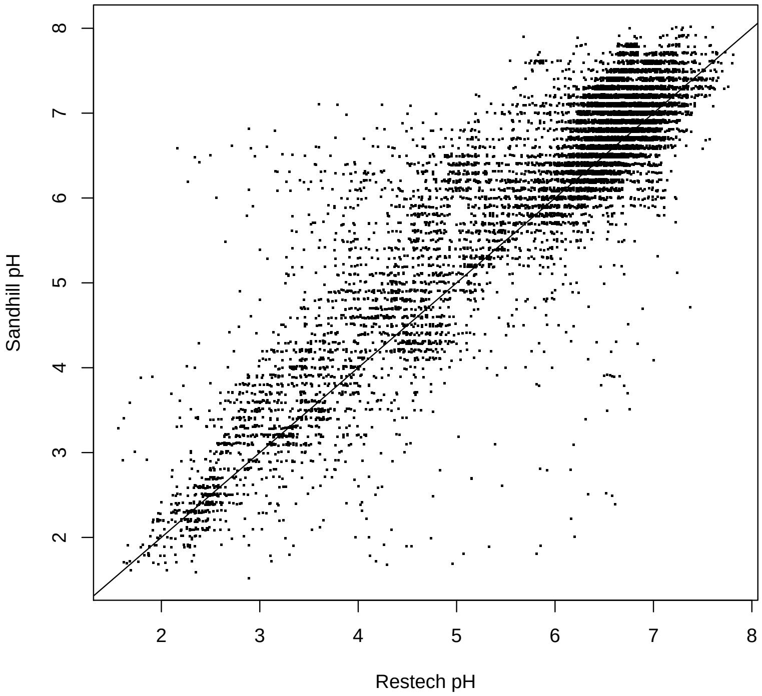


Figure 6.3: *Scatter plot of Restech and Sandhill pH readings. A 1:1 line is included to indicate "perfect" agreement between the two devices.*

- Incorrect approach would report a high correlation ($r = 0.90$) and conclude good agreement

- Problems with the correlation approach

  1. r measures the degree of linear association between two variables, not the agreement. If,for example, the Sandhill consistently gave pH values that were 0.5 unit higher than the Restech, we could still have high correlation, but poor agreement between the two devices. We can have high correlation if the two devices lie closely to any line, not just a 1:1 line that indicates perfect agreement.
  2. A change in scale does not affect correlation, but does influence agreement. For example, if the Sandhill always registered 2 times larger than the Restech, we would have perfect correlation but the agreement would get progressively worse for larger values of pH.
  3. Correlation depends on the range of the data so that larger ranges lead to larger correlations. This can lead to vary strange interpretations

	r	ρ
all data	0.90	0.73
avg pH ≤ 4	0.51	0.58
avg pH > 4	0.74	0.65

Table 6.1: *Pearson (r) and Spearman (ρ) correlations for Restech and Sandhill pH data. The correlation calculated using all of the data is larger than the correlation calculated using a retricted range of the data. However, it would be difficult to claim that the overall agreement is better than both the agreement when pH is less than 4 and when pH is greater than 4.*

  4. Tests of significance (testing if $r = 0$) are irrelevant to the question at hand, but often reported to demonstrate a significant association. The two devices are measuring the same quantity, so it would be shocking if we did not observe a highly significant p-value. A $p < .0001$ is not impressive. A regression analysis with a highly significant slope would be similarly unimpressive.
  5. Data can have high correlation, but poor agreement. There are many examples in the literature, but even in our analysis with $r = 0.90$, the correlation is high, but we will show that the agreement is not as good as the high correlation implies.

6.5.1 Bland-Altman Plots

EMS:36.4

- See Bland and Altman (1986, Lancet)

- Create plots of the difference in measurements on the y-axis versus the average value of the two devices on the x-axis

- If the two devices agree, the difference should be about zero

- The average of the two devices is our best estimate of the true, unknown (pH) value that is we are trying to measure

- Measurements will often vary in a systematic way over the range of measurement. By plotting the difference versus the average, we can visually determine if the difference changes over our estimate of the truth.

- Solid line indicated the mean, dashed lines are approximate 95% confidence intervals (assuming Normality)

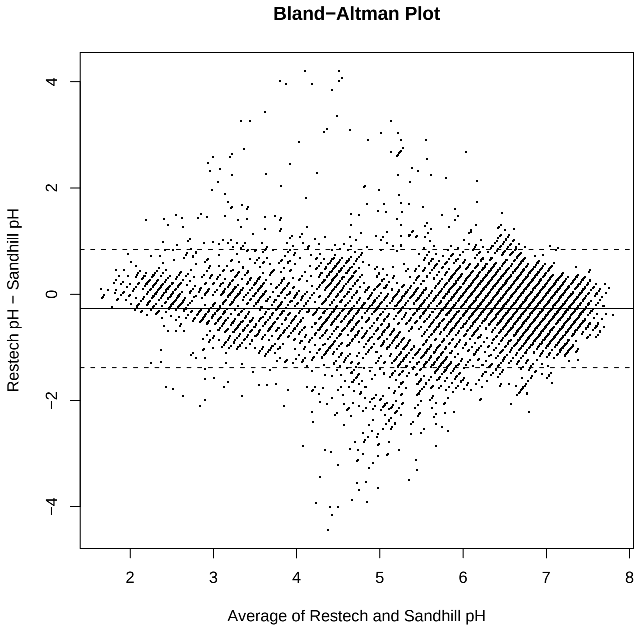


Figure 6.4: *Bland-Altman plot for the Restech and Sandhill pH data. The difference in pH mesaurements (Restech - Sandhill) is presented on the y-axis and the average of the two devices on the x-axis. We see poor agreement around pH values of 4-5*

- In our example, we will also consider differences in the two measurements over the time of day

- The added smooth curve is called a locally weighted scatterplot smooth (lowess)

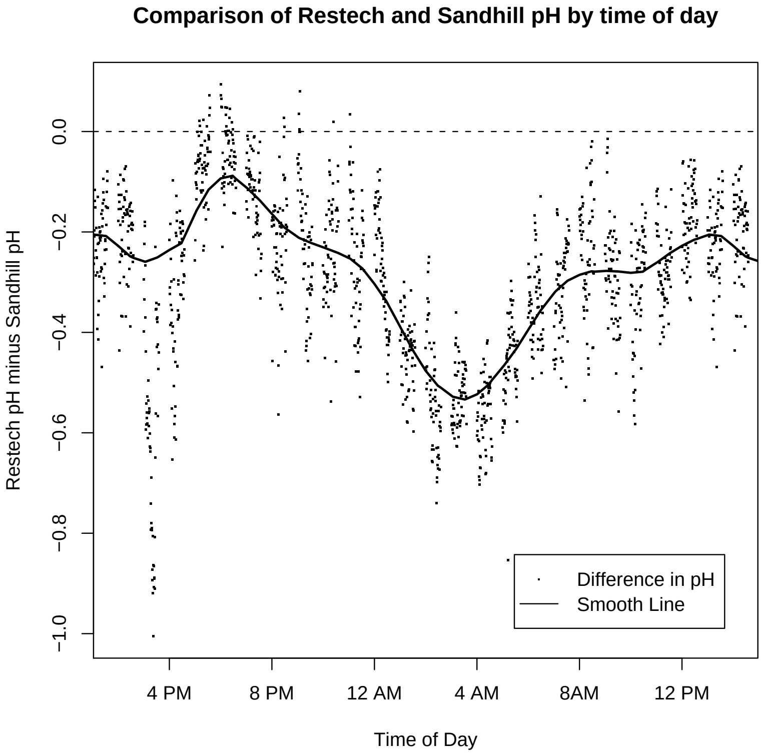


Figure 6.5: *Differene in pH measurements (Restech - Sandhill) by time of day. Is the difference modified by a subject being in a supine position rather than being upright?*

6.5.2 Using r to Compute Sample Size

- Without knowledge of population variances, etc., r can be useful for planning studies

- Choose n so that margin for error (half-width of C.L.) for r is acceptable

- Precision of r in estimating ρ is generally worst when $\rho = 0$

- This margin for error is shown in the figure below

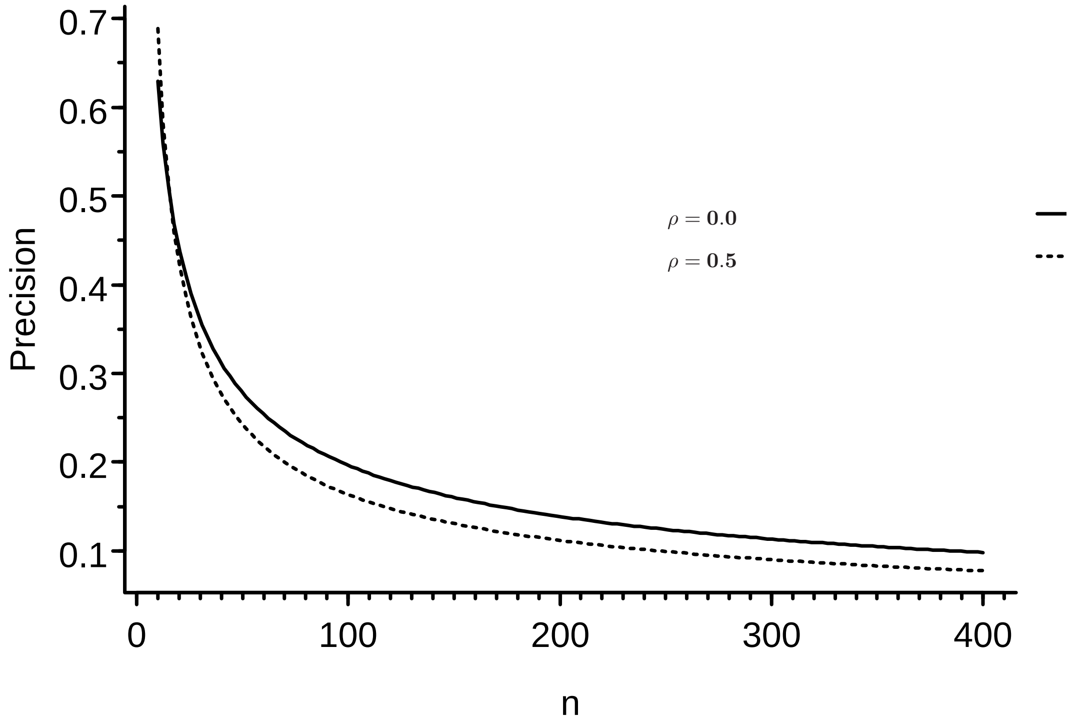


Figure 6.6: *Margin for error (length of longer side of asymmetric 0.95 confidence interval) for r in estimating ρ, when $\rho = 0$ (solid line) and $\rho = 0.5$ (dotted line). Calculations are based on Fisher's z transformation of r.*

6.5.3 Comparing Two r's

- Rarely appropriate

- Two r's can be the same even though slopes may differ

- Usually better to compare effects on a real scale (slopes)

Chapter 7

Simple and Multiple Regression Models

7.1 Purposes of Statistical Models

- Hypothesis testing
  - Test for no association (correlation) of a predictor (independent variable) and a response or dependent variable (unadjusted test) or test for no association of predictor and response after adjusting for the effects of other predictors

- Estimation
  - Estimate the shape and magnitude of the relationship between a single predictor (independent) variable and a response (dependent) variable

  - Estimate the effect on the response variable of changing a predictor from one value to another

- Prediction
  - Predicting response tendencies, e.g., long-term average response as a function of predictors

  - Predicting responses of individual subjects

7.2 Advantages of Modeling

Even when only testing H_0 a model based approach has advantages:

- Permutation and rank tests not as useful for estimation
- Cannot readily be extended to cluster sampling or repeated measurements
- Models generalize tests
  - 2-sample t-test, ANOVA $\rightarrow$ multiple linear regression
  - Wilcoxon, Kruskal-Wallis, Spearman $\rightarrow$ proportional odds ordinal logistic model
  - log-rank $\rightarrow$ Cox
- Models not only allow for multiplicity adjustment but for shrinkage of estimates
  - Statisticians comfortable with P-value adjustment but fail to recognize that the difference between the most different treatments is badly biased

Statistical estimation is usually model-based

- Relative effect of increasing cholesterol from 200 to 250 mg/dl on hazard of death, holding other risk factors constant
- Adjustment depends on how other risk factors relate to outcome

- Usually interested in adjusted (partial) effects, not unadjusted (marginal or crude) effects

7.3 Nonparametric Regression

- Estimate tendency (mean or median) of Y as a function of X
- Few assumptions
- Especially handy when there is a single X
- Plotted trend line may be the final result of the analysis
- Simplest smoother: moving average

X:	1	2	3	5	8
Y:	2.1	3.8	5.7	11.1	17.2

$$\begin{aligned}\hat{E}(Y|X=2) &= \frac{2.1+3.8+5.7}{3}\\ \hat{E}(Y|X=\frac{2+3+5}{3}) &= \frac{3.8+5.7+11.1}{3}\end{aligned}$$

  - overlap OK
  - problem in estimating $E(Y)$ at outer X-values
  - estimates very sensitive to bin width

- Moving linear regression far superior to moving avg. (moving flat line)

- Cleveland's moving linear regression smoother *loess* (locally weighted least squares) is the most popular smoother

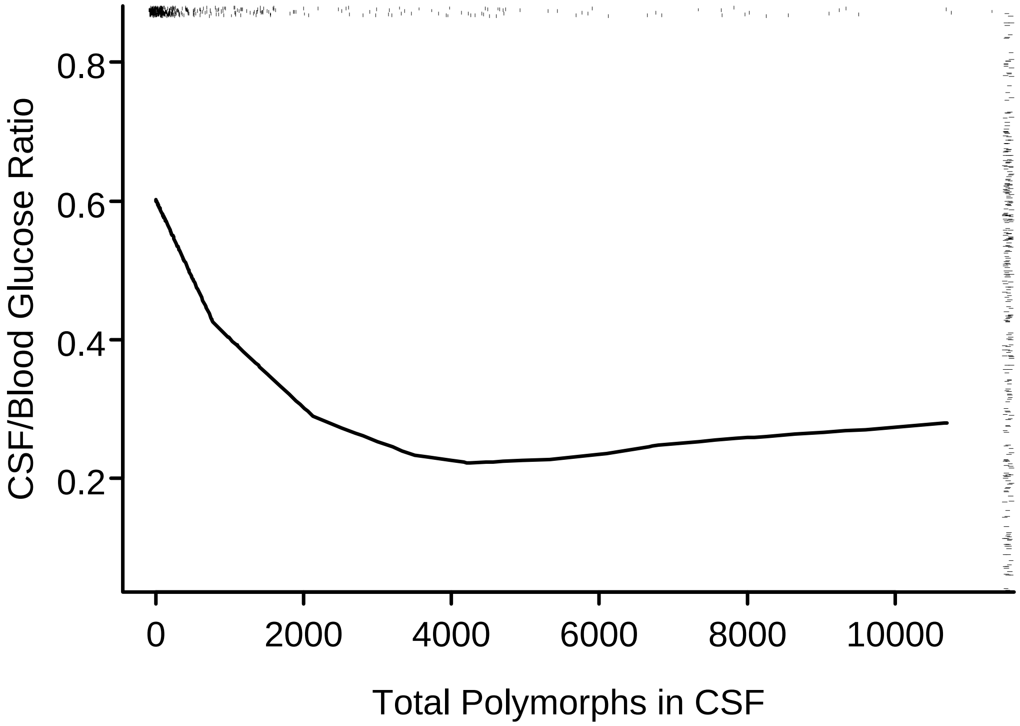


Figure 7.1: `loess` *nonparametric smoother relating CSF:blood glucose ratio to total CSF polymorph count in patients with either bacterial or viral meningitis. Rug plot on axes plots indicate raw data values.*

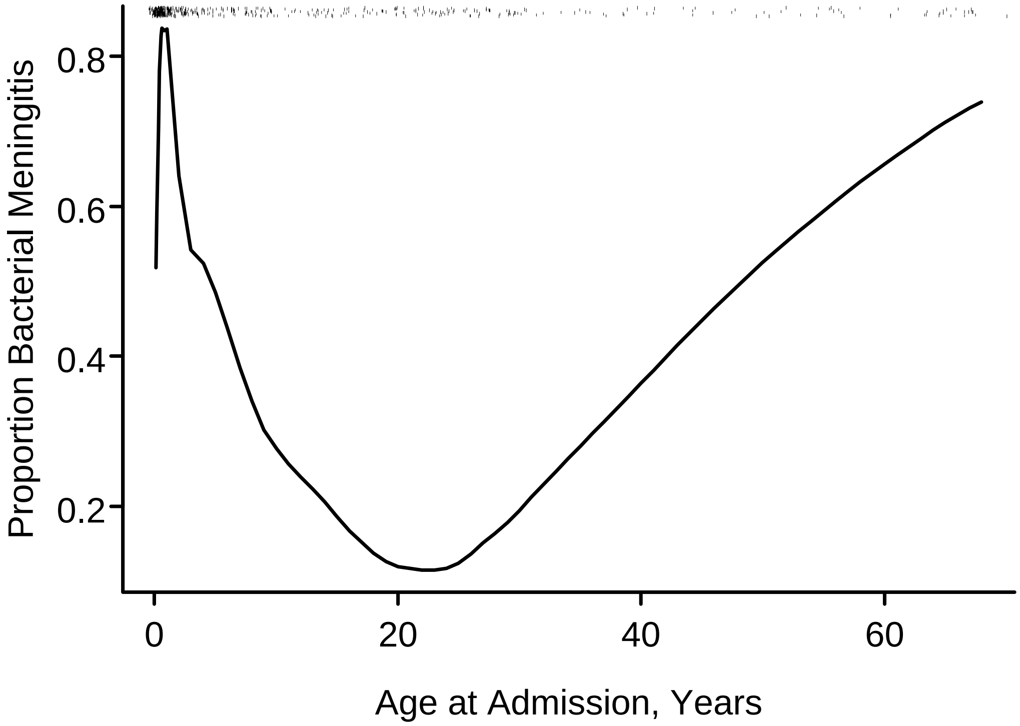


Figure 7.2: *"Super smoother" relating age to the probability of bacterial meningitis given a patient has bacterial or viral meningitis, with a rug plot showing the age distribution.*

7.4 Simple Linear Regression

11.1-6

7.4.1 Notation

- y : random variable representing response variable

- x : random variable representing independent variable (subject descriptor, predictor, covariable
  - conditioned upon

  - treating as constants, measured without error

- What does conditioning mean?
  - holding constant

  - subsetting on

  - slicing scatterplot vertically

- $E(y|x)$: population expected value or long-run average of y conditioned on the value of x
  Example: population average blood pressure for a 30-year old

- α : y-intercept

- β : slope of y on x ($\frac{\Delta y}{\Delta x}$)

Simple linear regression is used when

- Only two variables are of interest

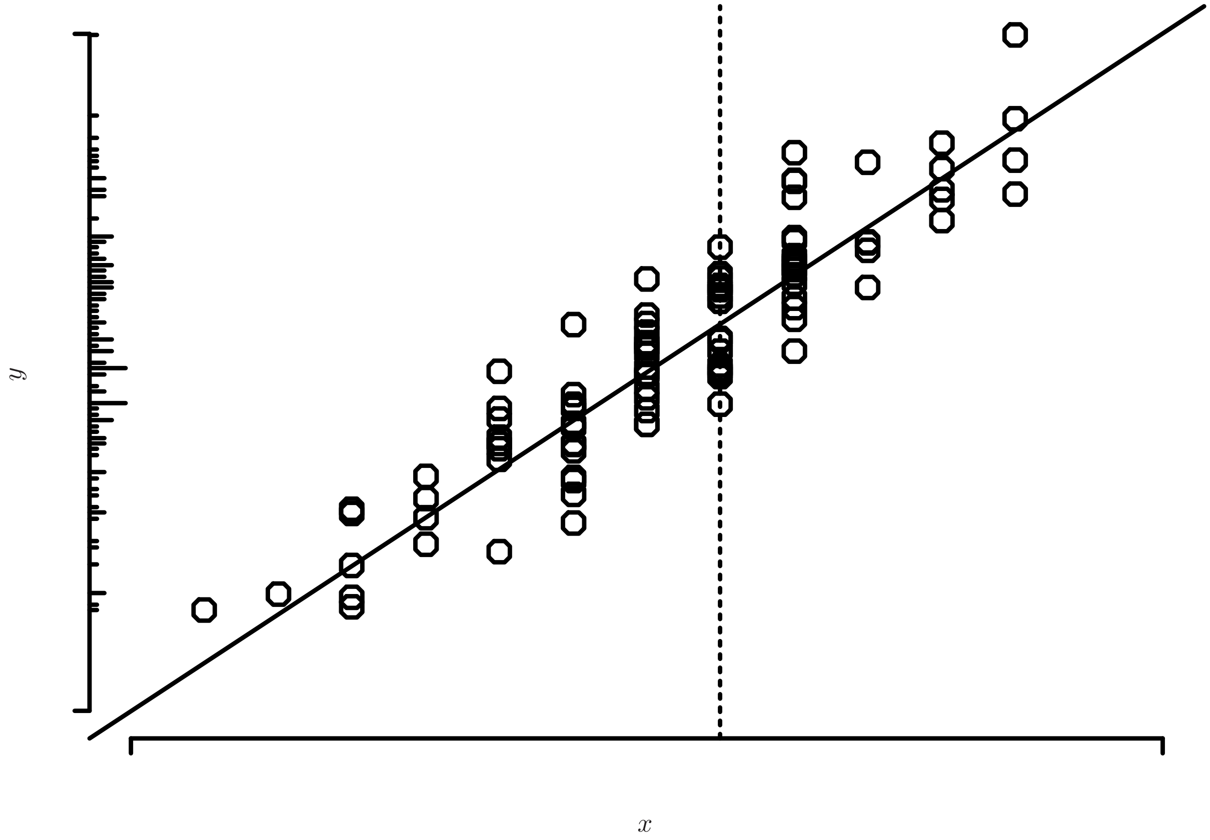


Figure 7.3: *Data from a sample of $n = 100$ points along with population linear regression line. The x variable is discrete. The conditional distribution of $y|x$ can be thought of as a vertical slice at x. The unconditional distribution of y is shown on the y-axis.*

- One variable is a response and one a predictor
- No adjustment is needed for confounding or other between-subject variation
- The investigator is interested in assessing the strength of the relationship between x and y in real data units, or in predicting y from x
- A linear relationship is assumed (why assume this? why not use nonparametric regression?)
- Not when one only needs to test for association (use Spearman's ρ rank correlation) or estimate a correlation index

7.4.2 Two Ways of Stating the Model

- $E(y|x) = \alpha + \beta x$
- $y = \alpha + \beta x + e$
  e is a random error (residual) representing variation between subjects in y even if x is constant, e.g. variation in blood pressure for patients of the same age

7.4.3 Assumptions, If Inference Needed

- Conditional on x, y is normal with mean $\alpha + \beta x$ and constant variance σ^2, **or:**
- e is normal with mean 0 and constant variance σ^2
- $E(y|x) = E(\alpha + \beta x + e) = \alpha + \beta x + E(e)$,
  $E(e) = 0$.

- Observations are independent

7.4.4 How Can α and β be Estimated?

- Need a criterion for what are good estimates
- **One** criterion is to choose values of the two parameters that minimize the sum of squared errors in predicting individual subject responses
- Let a, b be guesses for α, β
- Sample of size $n : (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- $SSE = \sum_{i=1}^{n}(y_i - a - bx_i)^2$
- Values that minimize SSE are *least squares estimates*
- These are obtained from

$$\begin{array}{cc} L_{xx} = \sum(x_i - \bar{x})^2 & L_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) \\ \hat{\beta} = b = \frac{L_{xy}}{L_{xx}} & \hat{\alpha} = a = \bar{y} - b\bar{x} \end{array}$$

- Note: A term from L_{xy} will be positive when x and y are concordant in terms of both being above their means or both being below their means.

7.4.5 Inference about Parameters

- Residual: $d = y - \hat{y}$

- d large if line was not the proper fit to the data or if there is large variability across subjects for the same x

- Beware of that many authors combine both components when using the terms *goodness of fit* and *lack of fit*

- Might be better to think of lack of fit as being due to a structural defect in the model (e.g., nonlinearity)

- $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
  $SSR = \sum(\hat{y}_i - \bar{y})^2$
  $SSE = \sum(y_i - \hat{y}_i)^2$
  $SST = SSR + SSE$
  $SSR = SST - SSE$

- SS increases in proportion to n

- Mean squares: normalized for for d.f.: $\frac{SS}{d.f.(SS)}$

- $MSR = SSR/p$, p = no. of parameters besides intercept (here, 1)
  $MSE = SSE/(n-p-1)$ (sample conditional variance of y)
  $MST = SST/(n-1)$ (sample unconditional variance of y)

- Brief review of ordinary ANOVA (analysis of variance):
  - Generalizes 2-sample t-test to > 2 groups
  - SSR is SS between treatment means
  - SSE is SS within treatments, summed over treatments

- ANOVA Table for Regression

Source	d.f.	SS	MS	F
Regression	p	SSR	$MSR = SSR/p$	MSR/MSE
Error	$n-p-1$	SSE	$MSE = SSE/(n-p-1)$	
Total	$n-1$	SST	$MST = SST/(n-1)$	

- Statistical evidence for large values of β can be summarized by $F = \frac{MSR}{MSE}$

- Has F distribution with p and $n-p-1$ d.f.

- Large values → $|\beta|$ large

7.4.6 Estimating σ, S.E. of $\hat{\beta}$; t-test

- $s^2_{y \cdot x} = \hat{\sigma}^2 = MSE = \widehat{Var}(y|x) = \widehat{Var}(e)$

- $\widehat{se}(b) = s_{y \cdot x}/L^{\frac{1}{2}}_{xx}$

- $t = b/\widehat{se}(b), n-p-1$ d.f.

- $t^2 \equiv F$ when $p = 1$

- $t_{n-2} \equiv \sqrt{F_{1,n-2}}$

- t identical to 2-sample t-test (x has two values)

- If x takes on only the values 0 and 1, b equals $\bar{y}$ when $x = 1$ minus $\bar{y}$ when $x = 0$

7.4.7 Interval Estimation

11.5

- 2-sided $1-\alpha$ CI for β: $b \pm t_{n-2,1-\alpha/2}\widehat{se}(b)$
- CI for *predictions* depend on what you want to predict even though $\hat{y}$ estimates both y [a] and $E(y|x)$
- Notation for these two goals: $\hat{y}$ and $\hat{E}(y|x)$
  - Predicting y with $\hat{y}$:
    $\widehat{s.e.}(\hat{y}) = s_{y\cdot x}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{L_{xx}}}$
    Note: This s.e. $\rightarrow s_{y\cdot x}$ as $n \rightarrow \infty$.
  - Predicting $\hat{E}(y|x)$ with $\hat{y}$:
    $\widehat{s.e.}(\hat{E}(y|x)) = s_{y\cdot x}\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{L_{xx}}}$ See footnote[b]
    Note: This s.e. shrinks to 0 as $n \rightarrow \infty$
- $1-\alpha$ 2-sided CI for either one:
  $\hat{y} \pm t_{n-p-1,1-\alpha/2}\widehat{s.e.}$
- Wide CI (large $\widehat{s.e.}$) due to:
  - small n
  - large σ^2
  - being far from the data center ($\bar{x}$)
- Example usages:

[a] With a normal distribution, the least dangerous guess for an individual y is the estimated mean of y.

[b] n here is the grand total number of observations because we are borrowing information about neighboring x-points, i.e., using interpolation. If we didn't assume anything and just computed mean y at each separate x, the standard error would instead by estimated by $s_{y\cdot x}\sqrt{\frac{1}{m}}$, where m is the number of original observations with x exactly equal to the x for which we are obtaining the prediction. The latter s.e. is much larger than the one from the linear model.

  - Is a child of age x smaller than predicted for her age?
    Use $s.e.(\hat{y})$

  - What is the best estimate of the population mean blood pressure for patients on treatment A?
    Use $s.e.(\hat{E}(y|x))$

- Example pointwise 0.95 confidence bands:

x	1	3	5	6	7	9	11
y:	5	10	70	58	85	89	135

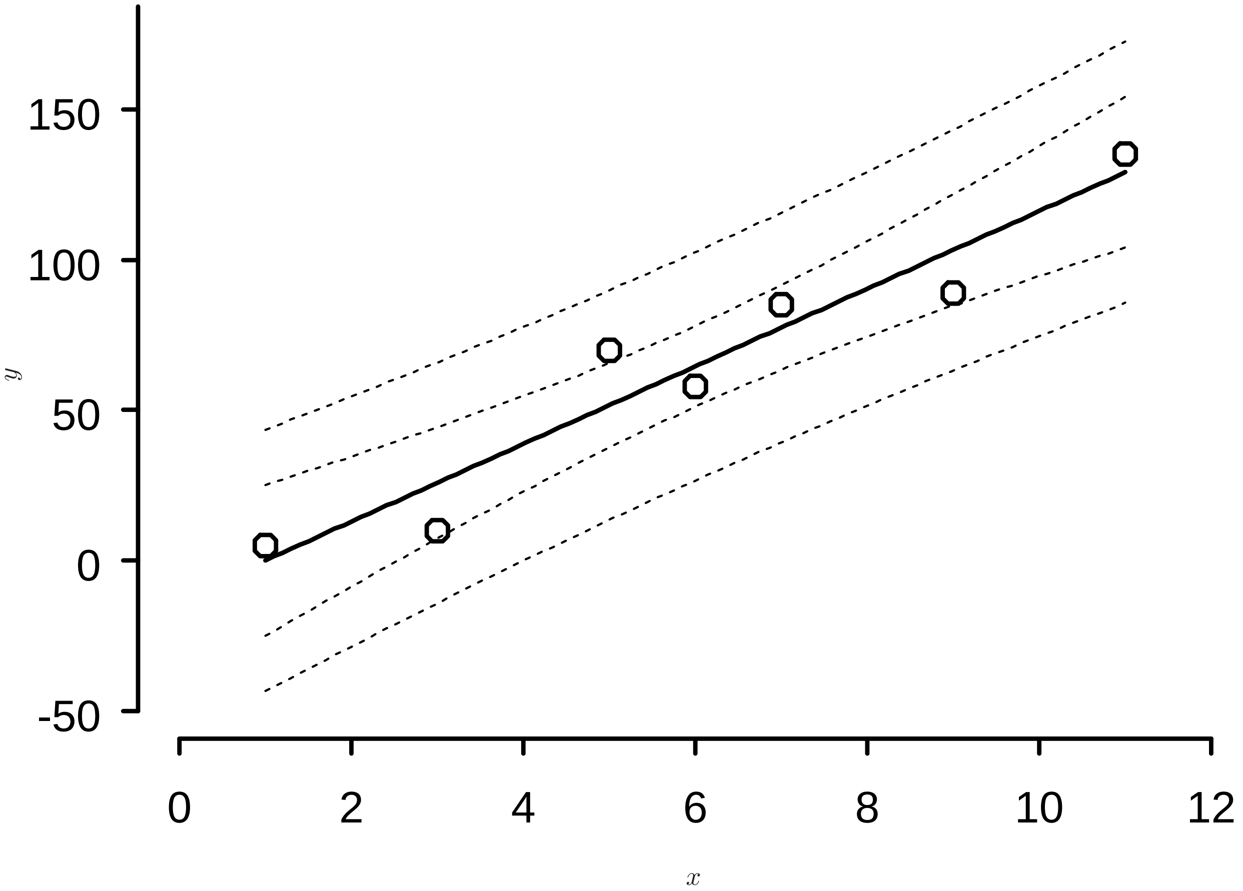


Figure 7.4: *Pointwise 0.95 confidence intervals for $\hat{y}$ (wider bands) and $\hat{E}(y|x)$ (narrower bands).*

7.4.8 Assessing Goodness of Fit

11.6

Assumptions:

1. Linearity
2. σ^2 is constant, independent of x
3. Observations (e's) are independent of each other
4. For proper statistical inference (CI, P-values), y (e) is normal conditional on x

Verifying some of the assumptions:

- In a scattergram the spread of y about the fitted line should be constant as x increases, and y vs. x should appear linear
- Easier to see this with a plot of $\hat{d} = y - \hat{y}$ vs. $\hat{y}$
- In this plot there are no systematic patterns (no trend in central tendency, no change in spread of points with x)
- Trend in central tendency indicates failure of linearity
- `qqnorm` plot of d

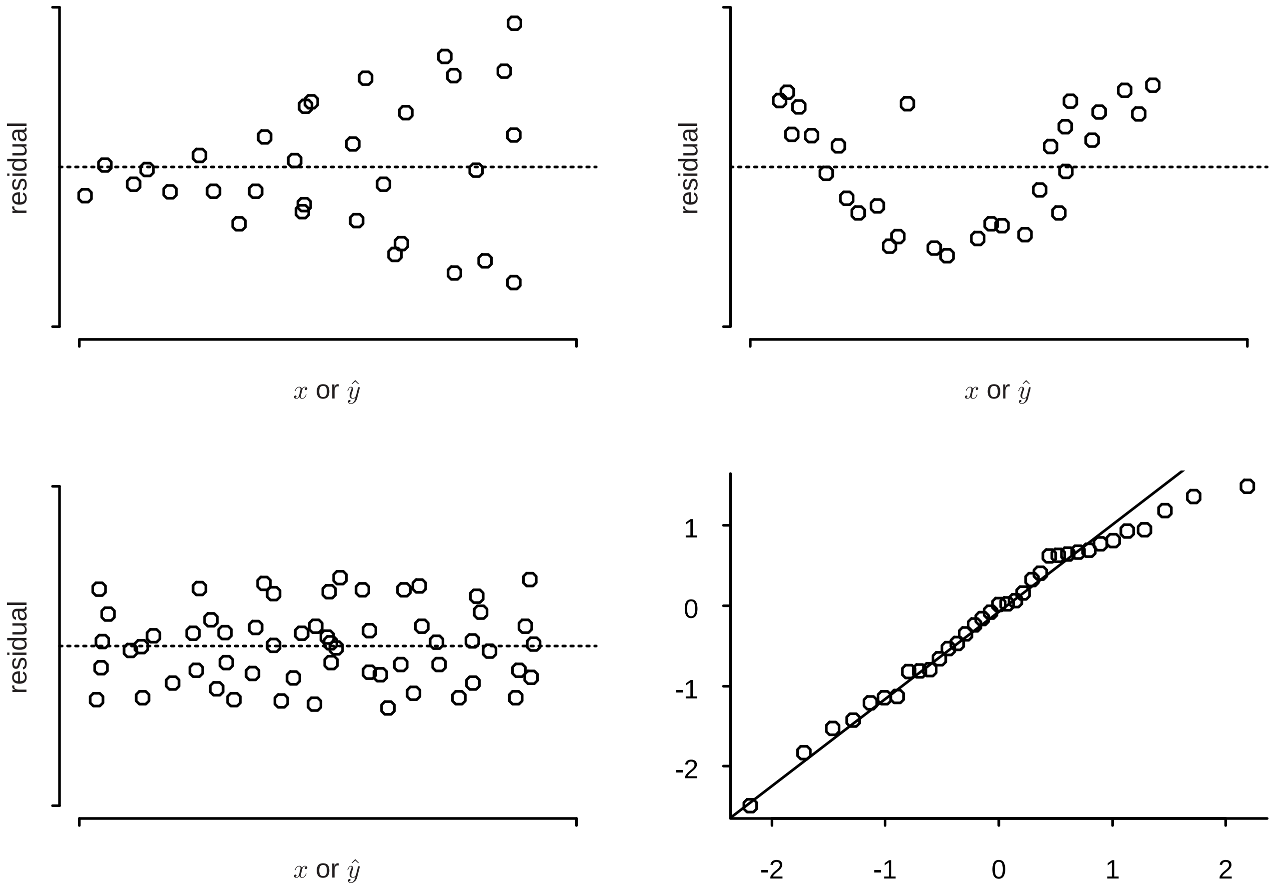


Figure 7.5: *Using residuals to check some of the assumptions of the simple linear regression model. Top left panel depicts non-constant σ^2, which might call for transforming y. Top right panel shows constant variance but the presence of a systemic trend which indicates failure of the linearity assumption. Bottom left panel shows the ideal situation of white noise (no trend, constant variance). Bottom right panel shows a $q-q$ plot that demonstrates approximate normality of residuals, for a sample of size $n = 35$. Horizontal reference lines are at zero, which is by definition the mean of all residuals.*

7.5 Multivariable Modeling

7.5.1 Examples of Uses of Predictive Multivariable Modeling

- Financial performance, consumer purchasing, loan pay-back
- Ecology
- Product life
- Employment discrimination
- Medicine, epidemiology, health services research
- Probability of diagnosis, time course of a disease
- Comparing non-randomized treatments
- Getting the correct estimate of relative effects in randomized studies requires covariable adjustment if model is nonlinear
  - Crude odds ratios biased towards 1.0 if sample heterogeneous
- Estimating absolute treatment effect (e.g., risk difference)
  - Use e.g. difference in two predicted probabilities
- Cost-effectiveness ratios
  - incremental cost / incremental *ABSOLUTE* benefit
  - most studies use avg. cost difference / avg. benefit, which may apply to no one

7.5.2 Planning for Modeling

- Chance that predictive model will be used [7]
- Response definition, follow-up
- Variable definitions
- Observer variability
- Missing data
- Preference for continuous variables
- Subjects
- Sites
- See [5]

Iezzoni [2] lists these dimensions to capture, for patient outcome studies:

1. age
2. sex
3. acute clinical stability
4. principal diagnosis
5. severity of principal diagnosis
6. extent and severity of comorbidities
7. physical functional status

8. psychological, cognitive, and psychosocial functioning
9. cultural, ethnic, and socioeconomic attributes and behaviors
10. health status and quality of life
11. patient attitudes and preferences for outcomes

7.5.3 Choice of the Model

- In biostatistics and epidemiology we usually choose model empirically
- Model must use data efficiently
- Should model overall structure (e.g., acute vs. chronic)
- Robust models are better
- Should have correct mathematical structure (e.g., constraints on probabilities)

7.6 Multiple Linear Regression

EMS 11

7.6.1 The Model and How Parameters are Estimated

- p independent variables $x_1, x_2, \ldots, x_p$
- Examples: multiple risk factors, treatment plus patient descriptors when adjusting for non-randomized treatment selection in an observational study; a set of controlled or uncontrolled factors in an experimental study; indicators of multiple experimental manipulations performed simultaneously

- Each variable has its own effect (slope) representing *partial effects*: effect of increasing a variable by one unit, holding all others constant

- Initially assume that the different variables act in an additive fashion

- Assume the variables act linearly against y

- Model: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e$

- Or: $E(y|x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

- For two x-variables: $y = \alpha + \beta_1 x_1 + \beta_2 x_2$

- Estimated equation: $\hat{y} = a + b_1 x_1 + b_2 x_2$

- Least squares criterion for fitting the model (estimating the parameters): $SSE = \sum_{i=1}^{n}[y - (a + b_1 x_1 + b_2 x_2)]^2$

- Solve for a, b_1, b_2 to minimize SSE

- When $p > 1$, least squares estimates require complex formulas; still all of the coefficient estimates are weighted combinations of the y's, $\sum w_i y_i$[c].

7.6.2 Interpretation of Parameters

- Regression coefficients are (b) are commonly called *partial regression coefficients*: effects of each variable holding all other variables in the model constant

- Examples of partial effects:

[c] When $p = 1$, the w_i for estimating β are $\frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$

  - model containing x_1=age (years) and x_2=sex (0=male 1=female) Coefficient of age (β_1) is the change in the mean of y for males when age increases by 1 year. It is also the change in y per unit increase in age for females. β_2 is the female minus male mean difference in y for two subjects of the same age.

  - $E(y|x_1, x_2) = \alpha + \beta_1 x_1$ for males, $\alpha + \beta_1 x_1 + \beta_2 = (\alpha + \beta_2) + \beta_1 x_1$ for females [the sex effect is a shift effect or change in y-intercept]

  - model with age and systolic blood pressure measured when the study begins Coefficient of blood pressure is the change in mean y when blood pressure increases by 1mmHg for subjects of the same age

- What is meant by changing a variable?

  - We usually really mean a comparison of two subjects with different blood pressures

  - Or we can envision what would be the expected response had *this* subject's blood pressure been 1mmHg greater at the outset[d]

  - We are not speaking of longitudinal changes in a single person's blood pressure

  - We can use subtraction to get the adjusted (partial) effect of a variable, e.g., $E(y|x_1, x_2) - \beta_2 x_2 = \alpha + \beta_1 x_1$

- Example: $\hat{y} = 37 + .01 \times \text{weight} + 0.5 \times \text{cigarettes smoked per day}$

  - .01 is the estimate of average increase y across subjects when weight is increased by 1lb. if cigarette smoking is unchanged

[d] This setup is the basis for randomized controlled trials and randomized animal experiments. Drug effects may be estimated with between-patient group differences under a statistical model.

  - 0.5 is the estimate of the average increase in y across subjects per additional cigarette smoked per day if weight does not change
  - 37 is the estimated mean of y for a subject of zero weight who does not smoke

- Comparing regression coefficients:
  - Can't compare directly because of different units of measurement. Coefficients in units of $\frac{y}{x}$.
  - Standardizing by standard deviations: not recommended. Standard deviations are not magic summaries of scale and they give the wrong answer when an x is categorical (e.g., sex).

7.6.3 What are Degrees of Freedom

For a model : the total number of parameters not counting intercept(s)

For a hypothesis test : the number of parameters that are hypothesized to equal specified constants. The constants specified are usually zeros (for *null* hypotheses) but this is not always the case. Some tests involve combinations of multiple parameters but test this combination against a single constant; the d.f. in this case is still one. Example: $H_0 : \beta_3 = \beta_4$ is the same as $H_0 : \beta_3 - \beta_4 = 0$ and is a 1 d.f. test because it tests one parameter ($\beta_3 - \beta_4$) against a constant (0).

These are **numerator d.f.** in the sense of the F-test in multiple linear regression. The F-test also entails a second kind of d.f., the **denominator** or **error** d.f., $n - p - 1$, where p is the number of parameters aside from the intercept. The error d.f. is the denominator of the estimator for σ^2 that is used to unbias the estimator, penalizing for having estimated $p + 1$ parameters by minimizing the sum of squared errors used to estimate σ^2 itself. You can think of the error d.f. as the sample size penalized for the number of parameters estimated, or as a measure of the information base used to fit the model.

7.6.4 Hypthesis Testing

EMS 11.2

Testing Total Association (Global Null Hypotheses)

- ANOVA table is same as before for general p
- $F_{p,n-p-1}$ tests $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$
- This is a test of *total association*, i.e., a test of whether *any* of the predictors is associated with y
- To assess total association we accumulate partial effects of all variables in the model *even though* we are testing if *any* of the partial effects is nonzero
- H_a : at least one of the β's is non-zero. **Note**: This does not mean that all of the x variables are associated with y.
- Weight and smoking example: H_0 tests the null hypothesis that neither weight nor smoking is associated with y. H_a is that at least one of the two variables is associated with y. The other may or may not have a non-zero β.
- Test of total association does not test whether cigarette smoking is related to y holding weight constant.
- SSR can be called the model SS

Testing Partial Effects

- $H_0 : \beta_1 = 0$ is a test for the effect of x_1 on y holding x_2 and any other x's constant
- Note that β_2 is *not* part of the null or alternative hypothesis; we assume that

we have adjusted for *whatever* effect x_2 has, *if any*

- One way to test β_1 is to use a t-test: $t_{n-p-1} = \frac{b_1}{\widehat{s.e.}(b_1)}$
- In multiple regression it is difficult to compute standard errors so we use a computer
- These standard errors, like the one-variable case, decrease when
  - $n \uparrow$
  - variance of the variable being tested $\uparrow$
  - σ^2 (residual y-variance) $\downarrow$
- Another way to get partial tests: the F test
  - Gives identical 2-tailed P-value to t test when one x being tested
    $t^2 \equiv$ partial F
  - Allows testing for > 1 variable
  - Example: is either systolic or diastolic blood pressure (or both) associated with the time until a stroke, holding weight constant
- To get a partial F define partial SS
- Partial SS is the change in SS when the variables **being tested** are dropped from the model and the model is re-fitted
- A general principle in regression models: a set of variables can be tested for their combined partial effects by removing that set of variables from the model and measuring the harm ($\uparrow SSE$) done to the model

- Let $full$ refer to computed values from the full model including all variables; $reduced$ denotes a reduced model containing only the adjustment variables and not the variables being tested
- Dropping variables $\uparrow SSE, \downarrow SSR$ unless the dropped variables had exactly zero slope estimates in the full model (which never happens)
- $SSE_{reduced} - SSE_{full} = SSR_{full} - SSR_{reduced}$
  Numberator of F test can use either SSE or SSR
- Form of partial F-test: change in SS when dropping the variables of interest divided by change in d.f., then divided by MSE;
  MSE is chosen as that which best estimates σ^2, namely the MSE from the full model
- Full model has p slopes; suppose we want to test q of the slopes

$$\begin{aligned} F_{q,n-p-1} &= \frac{(SSE_{reduced} - SSE_{full})/q}{MSE} \\ &= \frac{(SSR_{full} - SSR_{reduced})/q}{MSE} \end{aligned}$$

7.6.5 Assessing Goodness of Fit

EMS 12

Assumptions:

- Linearity of each predictor against y holding others constant
- σ^2 is constant, independent of x
- Observations (e's) are independent of each other

- For proper statistical inference (CI, P-values), y (e) is normal conditional on x

- x's act additively; effect of x_j does not depend on the other x's (**But** note that the x's may be correlated with each other without affecting what we are doing.)

Verifying some of the assumptions:

1. When $p = 2$, x_1 is continuous, and x_2 is binary, the pattern of y vs. x_1, with points identified by x_2, is two straight, parallel lines. β_2 is the slope of y vs. x_2 holding x_1 constant, which is just the difference in means for $x_2 = 1$ vs. $x_2 = 0$ as $\Delta x_2 = 1$ in this simple case.
2. In a residual plot ($d = y - \hat{y}$ vs. $\hat{y}$) there are no systematic patterns (no trend in central tendency, no change in spread of points with $\hat{y}$). The same is true if one plots d vs. any of the x's (these are more stringent assessments). If x_2 is binary box plots of d stratified by x_2 are effective.
3. Partial residual plots reveal the partial (adjusted) relationship between a chosen x_j and y, controlling for all other $x_i, i \neq j$, without assuming linearity for x_j. In these plots, the following quantities appear on the axes:

   y **axis:** residuals from predicting y from all predictors except x_j

   x **axis:** residuals from predicting x_j from all predictors except x_j (y is ignored)

   Partial residual plots ask how does what we can't predict about y without knowing x_j depend on what we can't predict about x_j from the other x's.

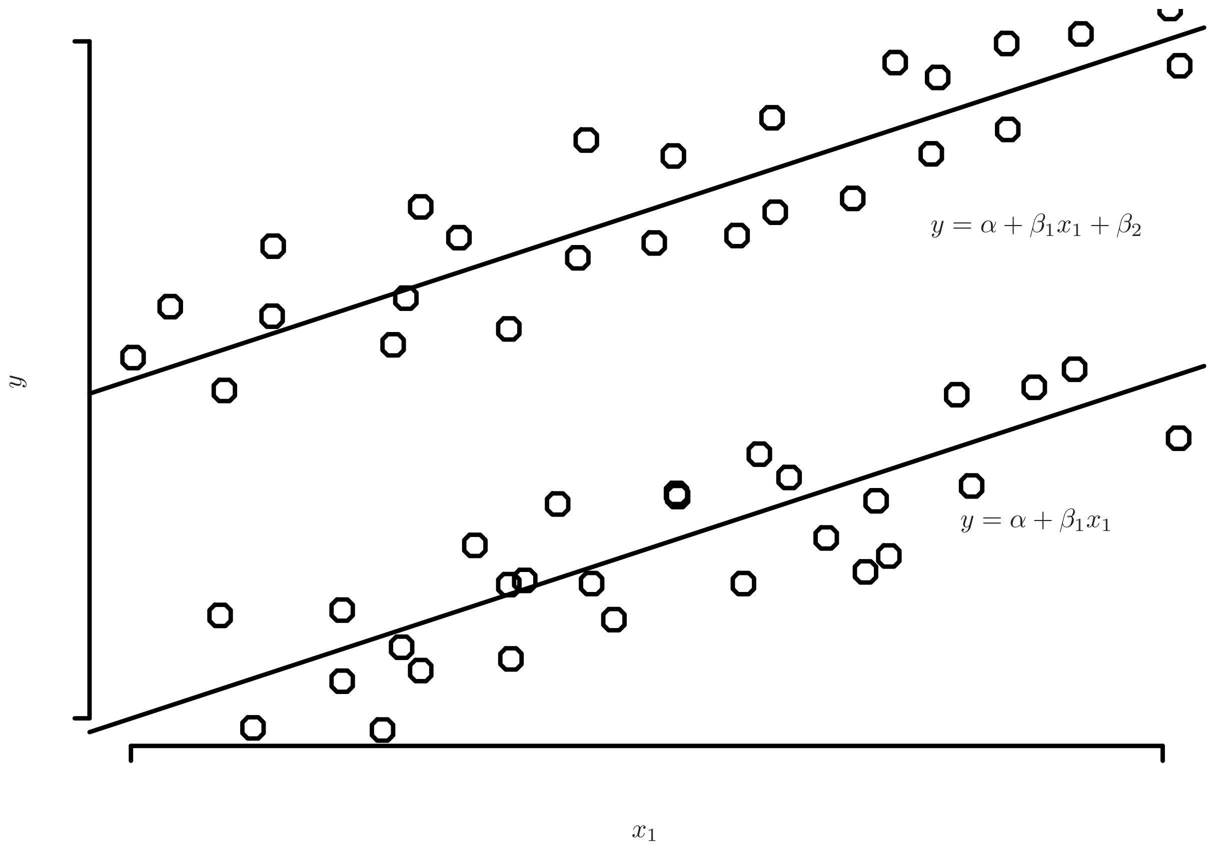


Figure 7.6: *Data satisfying all the assumptions of simple multiple linear regression in two predictors. Note equal spread of points around the population regression lines for the $x_2 = 1$ and $x_2 = 0$ groups (upper and lower lines, respectively) and the equal spread across x_1. The $x_2 = 1$ group has a new intercept, $\alpha + \beta_2$, as the x_2 effect is β_2.*

7.7 Case Study: Lead Exposure and Neuro-Psychological Function

7.7.1 Dummy Variable for Two-Level Categorical Predictors

- Categories of predictor: A, B (for example)

- First category = reference cell, gets a zero

- Second category gets a 1.0

- Formal definition of dummy variable: $x = I[category = B]$
  $I[w] = 1$ if w is true, 0 otherwise

- $\alpha + \beta x = \alpha + \beta I[category = B]$ =
  α for category A subjects
  $\alpha + \beta$ for category B subjects
  β = mean difference ($B - A$)

7.7.2 Two-Sample t-test vs. Simple Linear Regression

- They are equivalent in every sense:
  - P-value

  - Estimates and C.L.s after rephrasing the model

  - Assumptions (equal variance assumption of two groups in t-test is the same as constant variance of $y|x$ for every x)

- $a = \bar{Y}_A$
  $b = \bar{Y}_B - \bar{Y}_A$

- $\widehat{s.e.}(b) = \widehat{s.e.}(\bar{Y}_B - \bar{Y}_A)$

7.7.3 Analysis of Covariance

- Multiple regression can extend the t-test
  - More than 2 groups (multiple dummy variables can do multiple-group ANOVA)

  - Allow for categorical or continuous adjustment variables (covariates, co-variables)

- Model: $MAXFWT = \alpha + \beta_1 age + \beta_2 sex + e$

- Rosner coded $sex = 1, 2$ for male, female
  Does not affect interpretation of β_2 but makes interpretation of α more tricky (mean $MAXFWT$ when $age = 0$ and $sex = 0$ which is impossible by this coding.

- Better coding would have been $sex = 0, 1$ for male, female
  - α = mean $MAXFWT$ for a zero year-old male

  - β_1 = increase in mean $MAXFWT$ per 1-year increase in age

  - $\beta_2 =$ mean $MAXFWT$ for females minus mean $MAXFWT$ for males, holding age constant

- Model: $MAXFWT = \alpha + \beta_1 CSCN2 + \beta_2 age + \beta_3 sex + e$
  $CSCN2$ = 1 for exposed, 0 for unexposed

- β_1 = mean $MAXFWT$ for exposed minus mean for unexposed, holding age and sex constant

- Pay attention to Rosner's

  - t and F statistics and what they test
  - Figure 11.28 for checking for trend and equal variability of residuals (don't worry about standardizing residuals)

7.8 The Correlation Coefficient Revisited

Pearson product-moment linear correlation coefficient:

$$\begin{aligned} r &= \frac{Lxy}{\sqrt{L_{xx}L_{yy}}} \\ &= \frac{s_{xy}}{s_x s_y} \\ &= b\sqrt{\frac{L_{xx}}{L_{yy}}} \end{aligned}$$

- r is unitless
- r estimates the population correlation coefficient ρ (not to be confused with Spearman ρ rank correlation coefficient)
- $-1 \leq r \leq 1$
- $r = -1$: perfect negative correlation
- $r = 1$: perfect positive correlation
- $r = 0$: no correlation (no association)
- $t - test$ for r is identical to t-test for b

- r^2 is the proportion of variation in y explained by conditioning on x

- $(n-2)\frac{r^2}{1-r^2} = F_{1,n-2} = \frac{MSR}{MSE}$

- For multiple regression in general we use R^2 to denote the fraction of variation in y explained jointly by all the x's (variation in y explained by the whole model)

- $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1$ minus fraction of unexplained variation

- R^2 is called the *coefficient of determination*

- R^2 is between 0 and 1
  - 0 when $\hat{y}_i = \bar{y}$ for all i; $SSE = SST$
  - 1 when $\hat{y}_i = y_i$ for all i; SSE=0

- $R^2 \equiv r^2$ in the one-predictor case

7.9 Using Regression for ANOVA

EMS 9

7.9.1 Dummy Variables

`Lead Exposure Group`:

control : normal in both 1972 and 1973

currently exposed : elevated serum lead level in 1973, normal in 1972

previously exposed : elevated lead in 1972, normal in 1973

- Requires two dummy variables (and 2 d.f.) to perfectly describe 3 categories

- $x_1 = I[\text{currently exposed}]$
- $x_2 = I[\text{previously exposed}]$
- Reference cell is control
- Model:

$$\begin{aligned} E(y|exposure) &= \alpha + \beta_1 x_1 + \beta_2 x_2 \\ &= \alpha, \text{controls} \\ &= \alpha + \beta_1, \text{currently exposed} \\ &= \alpha + \beta_2, \text{previously exposed} \end{aligned}$$

  α : mean `maxfwt` for controls
  β_1 : mean `maxfwt` for currently exposed minus mean for controls
  β_2 : mean `maxfwt` for previously exposed minus mean for controls
  $\beta_2 - \beta_1$: mean for previously exposed minus mean for currently exposed

- In general requires $k - 1$ dummies to describe k categories
- For testing or prediction, choice of reference cell is irrelevant
- Does matter for interpreting individual coefficients
- Modern statistical programs automatically generate dummy variables from categorical or character predictors[e]
- In S never generate dummy variables yourself; just tell the functions you are using the name of the categorical predictor

[e] In S dummies are generated automatically any time a `factor` or `category` variable is in the model. For SAS you must list such variables in a `CLASS` statement.

7.9.2 Obtaining ANOVA with Multiple Regression

- Estimate α, β_j using standard least squares
- F-test for overall regression is exactly F for ANOVA
- In ANOVA, SSR is call sum of squares between treatments
- SSE is called sum of squares within treatments
- Don't need to learn formulas specifically for ANOVA

7.9.3 One-Way Analysis of Covariance

- Just add other variables (covariates) to the model
- Example: predictors age and treatment
  age is the covariate (adjustment variable)
- Global F test tests the global null hypothesis that neither age nor treatment is associated with response
- To test the adjusted treatment effect, use the partial F test for treatment based on the partial SS for treatment adjusted for age
- If treatment has only two categories, the partial t-test is an easier way to get the age-adjusted treatment test
- In R you can use

```
full ← ols(y  ∼ age + treat)
anova(full)     # actually gives you everything needed
reduced ← ols(y ∼ age)
```

```
anova(reduced)
# Subtract SSR or SSE from these two models to get treat effect
```

7.9.4 Two-Way ANOVA

- Two categorical variables as predictors
- Each variable is expanded into dummy variables
- One of the predictor variables may not be time or episode within subject; two-way ANOVA is often misused for analyzing repeated measurements within subject
- Example: 3 diet groups (NOR, SV, LV) and 2 sex groups
- $E(y|diet, sex) = \alpha + \beta_1 I[SV] + \beta_2 I[LV] + \beta_3 I[male]$
- Assumes effects of diet and sex are additive (separable) and not synergistic
- $\beta_1 = SV - NOR$ mean difference holding sex constant
  $\beta_3 = male - female$ effect holding diet constant
- Test of diet effect controlling for sex effect:
  $H_0 : \beta_1 = \beta_2 = 0$
  $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$
- This is a 2 d.f. partial F-test, best obtained by taking difference in SS between this full model and a model that excludes all diet terms.
- Test for significant difference in mean y for males vs. females, controlling for diet:
  $H_0 : \beta_3 = 0$

- For a model that has m categorical predictors (only), none of which interact, with numbers of categories given by $k_1, k_2, \ldots, k_m$, the total numerator regression d.f. is $\sum_{i=1}^{m}(k_i - 1)$

7.9.5 Two-way ANOVA and Interaction

Example: sex (F,M) and treatment (A,B)
Reference cells: F, A Model:

$$\begin{aligned} E(y|sex, treatment) &= \alpha + \beta_1 I[sex = M] \\ &+ \beta_2 I[treatment = B] + \beta_3 I[sex = M \cap treatment = B] \end{aligned}$$

Note that $I[sex = M \cap treatment = B] = I[sex = M] \times I[treatment = B]$.

α : mean y for female on treatment A (all variables at reference values)

β_1 : mean y for males minus mean for females, both on treatment A = sex effect holding treatment constant at A

β_2 : mean for female subjects on treatment B minus mean for females on treatment A = treatment effect holding sex constant at $female$

β_3 : $B - A$ treatment difference for males minus $B - A$ treatment difference for females
Same as $M - F$ difference for treatment B minus $M - F$ difference for treatment A

In this setting think of interaction as a “double difference”. To understand the parameters:

Group	$E(y\|Group)$
F A	α
M A	$\alpha + \beta_1$
F B	$\alpha + \beta_2$
M B	$\alpha + \beta_1 + \beta_2 + \beta_3$

Thus $MB - MA - [FB - FA] = \beta_2 + \beta_3 - \beta_2 = \beta_3$.

7.9.6 Interaction Between Categorical and Continuous Variables

This is how one allows the slope of a predictor to vary by categories of another variable. Example: separate slope for males and females:

$$\begin{aligned}
E(y|x) &= \alpha + \beta_1 age + \beta_2 I[sex = m] \\
&+ \beta_3 age \times I[sex = m] \\
E(y|age, sex = f) &= \alpha + \beta_1 age \\
E(y|age, sex = m) &= \alpha + \beta_1 age + \beta_2 + \beta_3 age \\
&= (\alpha + \beta_2) + (\beta_1 + \beta_3) age
\end{aligned}$$

α : mean y for zero year-old female

β_1 : slope of age for females

β_2 : mean y for males minus mean y for females, for zero year-olds

β_3 : increment in slope in going from females to males

Chapter 8

Multiple Groups

8.1 The k-Sample Problem

12.1

- When $k = 2$ we compare two means or medians, etc.

- When $k > 2$ we could do all possible pairwise 2-sample tests but this can be misleading and may ↑ type I error

- Advantageous to get a single statistic testing H_0: all groups have the same distribution (or at least the same central tendency)

8.2 Parametric ANOVA

12.2

- k samples each from a normal distribution

- Population means $\mu_1, \mu_2, \ldots, \mu_k$

- $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$
- H_1 : at least two of the population means differ
- Not placing more importance on any particular pair or combination although large samples get more weight in the analysis
- Assume that each of the k populations has the same σ
- If $k = 2$ ANOVA yields identical P-value as 2-tailed 2-sample t-test
- ANOVA uses an F statistic and is always 2-tailed
- F ratio is proportional to the sum of squared differences between each sample mean and the grand mean over samples, divided by the sum of squared differences between all raw values and the mean of the sample from which the raw value came
- This is the SSB/SSW (sum of squares between / sum of squares within)
- SSB is identical to regression sum of squares
  SSW is identical to sum of squared errors in regression
- $F = MSB/MSW$ where
  - MSB = mean square between = SSB/$(k-1), k-1 =$ “between group d.f.”
  - MSW = mean square within = SSW/$(n-k), n-k =$ “within group d.f.”
  - Evidence for different μs ↑ when differences in sample means (ignoring direction) are large in comparison to between-patient variation

- Can do ANOVA using multiple regression, using an intercept and $k-1$ “dummy” variables indicating group membership, so memorizing formulas specific to ANOVA is not needed
- Why is between group d.f.=$k-1$?
  - can pick any one group as reference group, e.g., group 1
  - H_0 is identical to $H_0: \mu_2 - \mu_1 = \mu_3 - \mu_1 = \ldots = \mu_k - \mu_1 = 0$
  - if $k-1$ differences in means are all zero, all means must be equal
  - since any unique $k-1$ differences define our goal, there is $k-1$ d.f. between groups for H_0

8.3 Why All These Distributions?

- Normal distribution is handy for approximating the distribution of z ratios (mean minus hypothesized value / standard error of mean) when n is large or σ is known
- If z is normal, z^2 has a χ^2_1 distribution
- If add k z^2 values the result has a χ^2_k distribution; useful
  - in larger than 2×2 contingency tables
  - in testing goodness of fit of a histogram against a theoretical distribution
  - when testing more than one regression coefficient in regression models not having a σ to estimate

- t distribution: when σ is estimated from the data; exact P-values if data from normal population
  Distribution indexed by d.f.: t_{df}; useful for
  - testing one mean against a constant
  - comparing 2 means
  - testing one regression coefficient in multiple linear regression
- t_{df}^2 has an F distribution
- F statistic can test
  - > 1 regression coefficient
  - > 2 groups
  - whether ratio of 2 variances=1.0 (this includes MSB/MSW)
- To do this F needs two different d.f.
  - numerator d.f.: how many unique differences being tested (like χ_k^2)
  - denominator d.f.
    - total sample size minus the number of means or regression coefficients and intercepts estimated from the data
    - is the denominator of the estimate of σ^2
    - also called the error or residual d.f.
- $t_{df}^2 = F_{1,df}$

- ANOVA results in $F_{k-1,df}$; d.f.=$N-k$ where $N=$ combined total sample size; cf. 2-sample t-test: d.f.=n_1+n_2-2

- Example: Ex. 12.4

  $$F = MSB/MSW = 58 \sim F_{4,1044}$$

  Use the cumulative distribution function calculator and plotter at `http://ebook.stat.ucla.edu/calculators/cdf` link from our web page (`surfstat` does not have the F distribution). The cumulative probability of getting an F statistic ≤ 58 with the above d.f. is 1.0000. We want Prob$(F \geq 58)$, so we get $P = 1 - 1 = 0$ to the accuracy of the calculator but report $P < 0.0001$.

8.4 Software and Data Layout

- Every general-purpose statistical package does ANOVA

- Small datasets are often entered using Excel

- Statistical packages expect a grouping variable, e.g., a column of treatment names or numbers; a column of response values for all treatments combines is also present

- If you enter different groups' responses in different spreadsheets or different columns within a spreadsheet, it is harder to analyze the data with a stat package

8.5 Comparing Specific Groups

12.4

- F test is for finding any differences but it does not reveal which groups are different

- Often it suffices to quote F and P, then to provide sample means (and their confidence intervals)

- Can obtain CLs for any specific difference using previously discussed 2-sample t-test, but this can result in inconsistent results due solely to sampling variability in estimating the standard error of the difference in means using only the two groups to estimate the common σ

- If assume that there is a common σ, estimate it using all the data to get a pooled s^2 Eq. 12.11

- $1-\alpha$ CL for $\mu_i - \mu_j$ is then

$$\bar{y}_i - \bar{y}_j \pm t_{n-k,1-\alpha/2} \times s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}},$$

  where n is the grand total sample size and there are respectively n_i and n_j observations in samples i and j

- Can test a specific $H_0 : \mu_i = \mu_j$ using similar calculations; Note that the d.f. for t comes from the grand sample size n, which ↑ power and ↓ width of CLs slightly

- Many people use more stringent α for individual tests when testing more than one of them (Section 8.9)
  - This is not as necessary when the overall F-test is significant

8.6 Kruskal-Wallis Test

- k-sample extension to the 2-sample Wilcoxon–Mann–Whitney rank-sum test

- Is very efficient when compared to parametric ANOVA even if data are from normal distributions

- Has same benefits as Wilcoxon (not harmed by outliers, etc.)
- Almost testing for equality of population medians
- In general, tests whether observations in one group tend to be larger than observations in another group (when consider randomly chosen pairs of subjects)
- Test statistic obtained by replacing all responses by their ranks across all subjects (ignoring group) and then doing an ANOVA on the ranks
- Compute F (many authors use a χ^2 approximation but F gives more accurate P-values)
- Look up against the F distribution with $k-1$ and $n-k$ d.f.
- Very accurate P-values except with very small samples
- Example: Ex. 12.21
  F statistic from ranks in Table 12.16: $F_{3,20} = 7.0289$
- Using the cumulative distribution calculator from the web page, the prob. of getting an F less impressive than this under H_0 is 0.9979
  P is $1 - 0.9979 = 0.0021$
- Compare with Rosner's $\chi^2_3 = 11.804$ from which $P = 0.008$ by `survstat` or one minus the CDF
- Evidence that not all of the 4 samples are from the same distribution
  - loosely speaking, evidence for differences in medians

  - better: some rabbits have larger anti-inflammatory effects when placed on different treatments in general

8.7 Two-Way ANOVA

12.6

- Ideal for a factorial design or observational study with 2 categorical grouping variables
- Example: 3 treatments are given to subjects and the researcher thinks that females and males will have different responses in general
  Six means: $\bar{Y}_{i,j}, i =$ treatment, $j =$ sex group
- Can test
  - whether there are treatment differences after accounting for sex effects
  - whether there are sex differences after accounting for treatment effects
  - whether the treatment effect is difference for females and males, if allow treatment × sex interaction to be in the model
- Suppose there are 2 treatments (A, B) and the 4 means are $\bar{Y}_{Af}, \bar{Y}_{Bf}, \bar{Y}_{Am}, \bar{Y}_{Bm}$, where f, m index the sex groups
- The various effects are estimated by
  - treatment effect: $\frac{(\bar{Y}_{Af}-\bar{Y}_{Bf})+(\bar{Y}_{Am}-\bar{Y}_{Bm})}{2}$
  - sex effect: $\frac{(\bar{Y}_{Af}-\bar{Y}_{Am})+(\bar{Y}_{Bf}-\bar{Y}_{Bm})}{2}$
  - treatment × sex interaction: $(\bar{Y}_{Af} - \bar{Y}_{Bf}) - (\bar{Y}_{Am} - \bar{Y}_{Bm}) = (\bar{Y}_{Af} - \bar{Y}_{Am}) - (\bar{Y}_{Bf} - \bar{Y}_{Bm})$

- Interactions are “double differences”
- Assessing whether treatment effect is same for m vs. f is the same as assessing whether the sex effect is the same for A vs. B
- **Note**: 2-way ANOVA is **not** appropriate when one of the categorical variables represents conditions applied to the same subjects, e.g. serially collected data within patient with time being one of the variables;
  2-way ANOVA assumes that all observations come from different subjects

8.8 Analysis of Covariance

12.5.3

- Generalizes two-way ANOVA
- Allows adjustment for continuous variables when comparing groups
- Can ↑ power and precision by reducing unexplained patient to patient variability (σ^2
- When Y is also measured at baseline, adjusting for the baseline version of Y can result in a major reduction in variance
- Fewer assumptions if adjust for baseline version of Y using ANCOVA instead of analyzing ($Y-$ baseline Y)
- Two-way ANOVA is a special case of ANCOVA where a categorical variable is the only adjustment variable (it is represented in the model by dummy variables)

8.9 Multiple Comparisons

12.4.3

- When hypotheses are prespecified and are few in number, don't need to correct P-values or α level in CLs for multiple comparisons

- Multiple comparison adjustments are needed with H_0 is effectively in the form
  - Is one of the 5 treatments effective when compared against control?
  - Of the 4 etiologies of disease in our patients, is the treatment effective in at least one of them?
  - Is the treatment effective in either diabetics, older patients, males, ..., etc.?
  - Diabetics had the greatest treatment effect empirically; the usual P-value for testing for treatment differences in diabetics was 0.03

- Recall that the probability that at least one event out of events $E_1, E_2, \ldots, E_m$ occurs is the sum of the probabilities if the events are mutually exclusive

- In general, the probability of at least one event is $\leq$ the sum of the probabilities of the individual events occurring

- Let the event be “reject H_0 when it is true”, i.e., making a type I error or false positive conclusion

- If test 5 hypotheses (e.g., 5 subgroup treatment effects) at the 0.05 level, the upper limit on the chance of finding one significant difference if there are no differences at all is $5 \times 0.05 = 0.25$

- This is called the *Bonferroni inequality*

- If we test each H_0 at the $\frac{\alpha}{5}$ level the chance of at least one false positive is no greater than α

- The chance of at least one false positive is the *experimentwise error probability* whereas the chance that a specific test is positive by chance alone is the *comparisonwise error probability*

- Instead of doing each test at the $\frac{\alpha}{m}$ level we can get a conservative adjusted P-value by multiplying an individual P-value by m[a]

- Whenever $m \times P > 1.0$ report $P = 1.0$

- There are many specialized and slightly less conservative multiple comparison adjustment procedures. Some more complex procedures are actually more conservative than Bonferroni.

- Statisticians generally have a poor understanding about the need to not only adjust P-values but to adjust point estimates also, when many estimates are made and only the impressive ones (by P) are discussed. In that case point estimates are badly biased away from the null value. For example, the BARI study analyzed around 20 subgroups and only found a difference in survival between PTCA and CABG in diabetics. The hazard ratio for CABG:PTCA estimated from this group is far too extreme.

[a] Make sure that m is the total number of hypotheses tested with the data, whether formally or informally.

Chapter 9

Statistical Inference Review

- Emphasize confidence limits, which can be computed from adjusted or unadjusted analyses, with or without taking into account multiple comparisons

- P-values can accompany CLs if formal hypothesis testing needed

- When possible construct P-values to be consistent with how CLs are computed

9.1 Types of Analyses

- Except for one-sample tests, all tests can be thought of as testing for an association between at least one variable with at least one other variable

- Testing for group differences is the same as testing for association between group and response

- Testing for association between two continuous variables can be done using correlation (especially for unadjusted analysis) or regression methods; in simple cases the two are equivalent

- Testing for association between group and outcome, when there are more than 2 groups which are not in some solid order[a] means comparing a summary of the response between k groups, sometimes in a pairwise fashion

9.2 Covariable-Unadjusted Analyses

Appropriate when

- Only interested in assessing the relationship between a single X and the response, or
- Treatments are randomized and there are no strong prognostic factors that are measureable
- Study is observational and variables capturing confounding are unavailable (place strong caveats in the paper)

9.2.1 Analyzing Paired Responses

Type of Response	Recommended Test	Most Frequent Test
binary	McNemar	McNemar
continuous	Wilcoxon signed-rank	paired t-test

[a] The dose of a drug or the severity of pain are examples of ordered variables.

9.2.2 Comparing Two Groups

Type of Response	Recommended Test	Most Frequent Test
binary	$2 \times 2 \chi^2$	χ^2, Fisher's exact test
ordinal	Wilcoxon 2-sample	Wilcoxon 2-sample
continuous	Wilcoxon 2-sample	2-sample t-test
time to event[a]	Cox model[b]	log-rank[c]

[a] The response variable may be right-censored, which happens if the subject ceased being followed before having the event. The value of the response variable, for example, for a subject followed 2 years without having the event is 2+.

[b] If the treatment is expected to have more early effect with the effect lessening over time, an accelerated failure time model such as the lognormal model is recommended.

[c] The log-rank is a special case of the Cox model. The Cox model provides slightly more accurate P-values than the χ^2 statistic from the log-rank test.

9.2.3 Comparing > 2 Groups

Type of Response	Recommended Test	Most Frequent Test
binary	$r \times 2 \chi^2$	χ^2, Fisher's exact test
ordinal	Kruskal-Wallis	Kruskal-Wallis
continuous	Kruskal-Wallis	ANOVA
time to event	Cox model	log-rank

9.2.4 Correlating Two Continuous Variables

Recommended: Spearman ρ
Most frequently seen: Pearson r

9.3 Covariable-Adjusted Analyses

- To adjust for imbalances in prognostic factors in an observational study or for strong patient heterogeneity in a randomized study

- Analysis of covariance is preferred over stratification, especially if continuous adjustment variables are present or there are many adjustment variables

- Continuous response: multiple linear regression with appropriate transformation of Y

- Binary response: binary logistic regression model

- Ordinal response: proportional odds ordinal logistic regression model

- Time to event response, possibly right-censored:
  - chronic disease: Cox proportional hazards model

  - acute disease: accelerated failure time model

Chapter 10

Measuring Change

10.1 Analysis of Paired Observations

- Frequently one makes multiple observations on same experimental unit
- Can’t analyze as if independent
- When two observations made on each unit (e.g., pre–post), it is common to summarize each pair using a measure of effect → analyze effects as if (unpaired) raw data
- Most common: simple difference, ratio, percent change
- Can’t take effect measure for granted
- Subjects having large initial values may have largest differences
- Subjects having very small initial values may have largest post/pre ratios

10.2 What's Wrong with Percent Change?

- Depends on point of reference — which term is used in the denominator?

- Example:
  Treatment A: 0.05 proportion having stroke
  Treatment B: 0.09 proportion having stroke
  Treatment A reduced proportion of stroke by 44%
  Treatment B increased proportion by 80%

- Two increases of 50% result in a total increase of 125%, not 100%

- Percent change (or ratio) not a symmetric measure

- Simple difference or log ratio are symmetric

10.3 Objective Method for Choosing Effect Measure

- Goal: Measure of effect should be as independent of baseline value as possible[a]

- Plot difference in pre and post values vs. the average of the pre and post values. If this shows no trend, the simple differences are adequate summaries of the effects, i.e., they are independent of initial measurements.

- If a systematic pattern is observed, consider repeating the previous step after taking logs of both the pre and post values. If this removes any systematic relationship between the average and the difference in logs, summarize the data using logs, i.e., take the effect measure as the log ratio.

[a] Because of regression to the mean, it may be impossible to make the measure of change truly independent of the initial value. A high initial value may be that way because of measurement error. The high value will cause the change to be less than it would have been had the initial value been measured without error. Plotting differences against averages rather than against initial values will help reduce the effect of regression to the mean.

- Other transformations may also need to be examined

Chapter 11

Modeling for Observational Treatment Comparisons

11.1 Propensity Score

- In observational studies comparing treatments, need to adjust for nonrandom treatment selection
- Number of confounding variables can be quite large
- May be too large to adjust for them using multiple regression, due to overfitting (may have more potential confounders than outcome events)
- Assume that all factors related to treatment selection that are prognostic are collected
- Use them in a flexible regression model to predict treatment actually received (e.g., logistic model allowing nonlinear effects)
- **Propensity score** (PS) = estimated probability of getting treatment B vs.

treatment A

- Use of the PS allows one to aggressively adjust for confounders by simulating a randomized trial

- Doing an adjusted analysis where the adjustment variable is the PS simultaneously adjusts for all the variables in the score

- If after adjusting for the score there were a residual imbalance for one of the variables, that would imply that the variable was not correctly modeled in the PS

- E.g.: after holding PS constant there are more subjects above age 70 in treatment B; means that age> 70 is still predictive of treatment received after adjusting for PS, or age> 70 was not modeled correctly.

11.2 Assessing Treatment Effect

- Eliminate patients in intervals of PS where there is no overlap between A and B

- Many researchers stratify the PS into quintiles, get treatment differences within the quintiles, and average these to get adjustment treatment effects

- Often results in imbalances in outer quintiles due to skewed distributions of PS there

- Can do a matched pairs analysis but depends on matching tolerance and many patients will be discarded when their case has already been matched

- Usually better to adjust for PS in a regression model

- Model: $Y = \text{treat} + \log \frac{PS}{1-PS} +$ nonlinear functions of $\log \frac{PS}{1-PS} +$ important prognostic variables

- Prognostic variables need to be in outcome (Y) model even though they are also in the PS, to account for patient heterogeneity (susceptibility bias)

- If outcome is binary and can afford to ignore prognostic variables, use nonparametric regression to relate PS to outcome separately in actual treatment A vs. B groups

- Plotting these two curves with PS on x-axis and looking at vertical distances between curves is an excellent way to adjust for PS continuously without assuming a model

11.3 Sensitivity Analysis

- For n patients in the analysis, generate n random values of a hypothetical unmeasured confounder U

- Constrain U so that the effect of U on the response Y is given by an adjusted odds ratio of OR_Y and so that U's distribution is unbalanced in group A vs. B to the tune of an odds ratio of OR_{treat}.

- Solve for how large OR_Y and OR_{treat} must be before the adjusted treatment effect reverses sign or changes in statistical significance

- The larger are OR_Y and OR_{treat} the less plausible it is that such an unmeasured confounder exists

Bibliography

[1] T. J. Cole. Sympercents: symmetric percentage differences on the 100 $\log_e$ scale simplify the presentation of log transformed data. *Stat Med*, 19:3109–3125, 2000.

KEY: col00sym
ANNOTATION: measuring change;quantifying change

[2] Lisa I. Iezzoni. Dimensions of risk. In Lisa I. Iezzoni, editor, *Risk Adjustment for Measuring Health Outcomes*, chapter 2, pages 29–118. Foundation of the American College of Healthcare Executives, Ann Arbor, MI, 1994.

KEY: iez94ris
ANNOTATION: dimensions of risk factors to include in mdoels

[3] Lee Kaiser. Adjusting for baseline: Change or percentage change? *Stat Med*, 8:1183–1190, 1989.

KEY: kai89
ANNOTATION: Research methods, measurement, Miscellaneous; measuring change; one-sample problem; percent change

[4] R. A. Kronmal. Spurious correlation and the fallacy of the ratio standard revisited. *J Roy Stat Soc A*, 156:379–392, 1993.

KEY: kro93spu
ANNOTATION: spurious correlation in using ratio variables even if all component variables of ratios are uncorrelated;measuring change;ratio;division of only the dependent variable by an independent variable can result in regression coefficient estimates for the other independent variables that result in inappropriate conclusions;use of a ratio as an independent variable can result in inadequate adjustment for component variables of the ratio;ratio variables

should only be used in a full model containing all the component variables;results of regression analyses incorporating ratios are not readily comparable across studies

[5] Andreas Laupacis, Nandita Sekar, and Ian G. Stiell. Clinical prediction rules: A review and suggested modifications of methodological standards. *JAMA*, 277:488–494, 1997.

KEY: lau97cli
ANNOTATION: teaching MDs;clinical prediction;multivariable modeling;ROC curves are not of interest to clinicians;debate about choosing a rule with sensitivity of 1.0;use of probabilities instead of classifications;reporting of statistical results

[6] J. S. Maritz. Models and the use of signed rank tests. *Stat Med*, 4:145–153, 1985.

KEY: mar85mod
ANNOTATION: measuring change;percent change;one-sample problem;signed rank test

[7] Brendan M. Reilly and Arthur T. Evans. Translating clinical research into clinical practice: Impact of using prediction rules to make decisions. *Ann Int Med*, 144:201–209, 2006.

KEY: rei06tra
ANNOTATION: clinical prediction rule;planning for modeling;impact analysis;example of decision aid being ignored or overruled making MD decisions worse;assumed utilities are constant across subjects by concluding that directives have more impact than predictions;Goldman-Cook clinical prediction rule in AMI

[8] L. Törnqvist, P. Vartia, and Y. O. Vartia. How should relative changes be measured? *Ann Math Stat*, 39:43–46, 1985.

KEY: tor85how
ANNOTATION: measuring change; percent change; one-sample problem