1. Introduction

Example of a medical experiment (major employers of statisticians are hospitals, pharmaceutical firms, medical research facilities ...)

- Two headache remedies are to be compared, to see which is more effective at alleviating the symptoms.

- Control (Drug 1) and New (Drug 2) - Drug type is the factor whose effect on the outcome is to be studied.

- **Outcome** variable: rate of blood flow to the brain, one hour after taking the drug.

- Other factors may well affect the outcome - gender of the patient (M or F), dosage level (LO or HI), etc.
• Single factor experiment: here the drug type will be the only factor in our statistical model.

  – Patients suffering from headaches volunteer for the study, and are randomly divided into two groups. One group receives Drug 1, the other receives Drug 2. Within each group the dosage levels are randomly assigned.

  – We hope that the randomization will control for the other factors, in that (on average) there should be approximately equal numbers of men and women in each group, equal numbers at each dosage level, etc.
• A statistical model: Suppose that $n$ patients receive Drug 1 ($i = 1$) and $n$ receive Drug 2 ($i = 2$). Define $y_{ij}$ to be the response of the $j^{th}$ subject ($j = 1, \ldots, n$) in the $i^{th}$ group. A possible model is

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

where the parameters are:

- $\mu$ = ‘overall mean’ (mean effect of the drugs, irrespective of which one is used), estimated by the overall average

$$\bar{y}_{..} = \frac{1}{2n} \sum_{i=1}^{2} \sum_{j=1}^{n} y_{ij},$$

- $\tau_i$ = ‘$i^{th}$ treatment effect’, estimated by

$$\hat{\tau}_i = \bar{y}_i - \bar{y}_{..},$$

where

$$\bar{y}_i = \frac{1}{n} \sum_{j=1}^{n} y_{ij},$$

and
\[ \epsilon_{ij} = \text{‘random error’ - measurement error, and effects of factors which are not included in the model, but perhaps should be.} \]

- We might instead control for the dosage level as well. We could obtain \(4n\) volunteers, split each group into four (randomly) of size \(n\) each and assign these to the four combinations Drug 1/LO dose, Drug 1/HI dose, Drug 2/LO dose, Drug 2/HI dose. This is a \(2^2\) factorial experiment: 2 factors - drug type, dosage - each at 2 levels, for a total of \(2 \times 2 = 2^2\) combinations of levels and factors.

- Now the statistical model becomes more involved - it contains parameters representing the effects of each of the factors, and possibly interactions between them (e.g., Drug 1 may only be more effective at HI dosages, Drug 2 at LO dosages).
• We could instead split the group into men and women (the 2 blocks) of subjects and run the entire experiment, as described above, within each block. We expect the outcomes to be more alike within a block than not, and so the blocking can reduce the variation of the estimates of the factor effects.

• There are other experimental designs which control for the various factors but require fewer patients - Latin squares, Graeco-Roman squares, repeated measures, etc.

• Designing an experiment properly takes a good deal of thought and planning. In this course we will be doing MUCH MORE than presenting formulas and plugging numbers into them.
• **Planning** phase of an experiment
  
  – What is to be measured? (i.e. what is the response variable?)
  
  – What are the influential factors?

• **Experimental design** phase
  
  – Control the known sources of variation (as in the factorial experiment described above, or through blocking)
  
  – Allow estimation of the size of the uncontrolled variation (e.g. by assigning subjects to drug type/dosage level blocks randomly, so as to average out a ‘gender’ effect).

• **Statistical analysis** phase
  
  – Make inferences on factors included in the statistical model
  
  – Suggest more appropriate models
Part II

SIMPLE COMPARATIVE EXPERIMENTS
2. Basic concepts; sampling distributions

• Example of an industrial experiment: compare two types of cement mortar, to see which results in a stronger bond. We’ll call these ‘modified’ and ‘unmodified’; more details in text. Data collection: 10 samples of the unmodified formulation were prepared, and 10 of the modified were prepared. They are prepared and their 20 bond strengths measured, in random order, by randomly assigned technicians, etc. (A ‘completely randomized’ design.)

• Data in text and on course web site:

```r
> mod
16.85 16.40 17.21 16.35 16.52
17.04 16.96 17.15 16.59 16.57
> unmod
17.50 17.63 18.25 18.00 17.86
17.75 18.22 17.90 17.96 18.15
```
• Summary statistics:

```r
> summary(mod)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
  16.35 16.53  16.72 16.76 17.02  17.21
> summary(unmod)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
  17.50 17.78  17.93 17.92 18.11  18.25
```
Graphical displays:

- `boxplot(mod, unmod)` yields the following boxplots (Q2=median, ‘hinges’ = approx. Q1, Q3, ‘whiskers’ go to the most extreme observations within 1.5 IQR of Q2).

Fig 2.1
Some basic notions from mathematical statistics

• $X$ a random variable, e.g. strength of a randomly chosen sample of unmodified mortar. (Think about how the randomness might arise.)

• $E[X]$ is the expected value, or mean, of $X$; also written $\mu_X$ (or just $\mu$). It can be calculated from the probability distribution of $X$; see §2.2 or review STAT 265 for details.

• If $f(X, Y)$ is a function of r.v.s $X$ and $Y$ such as $X^2$ or $XY$, then $f(X, Y)$ is another r.v.; call it $Z$ and then $E[f(X, Y)]$ is defined to be $E[Z]$.

• Basic tool is linearity:

$$E[aX + bY + c] = aE[X] + bE[Y] + c$$

for constants $a$, $b$ and $c$. 
• Example:

\[
E \left[ (X - \mu_X)^2 \right] = E \left[ X^2 - 2\mu_X X + \mu_X^2 \right] \\
= E \left[ X^2 \right] - 2\mu_X E [X] + \mu_X^2 \\
= E \left[ X^2 \right] - \mu_X^2.
\]

The quantity being studied here is the variance \(\sigma_X^2\). Its square root \(\sigma_X\) is the standard deviation. The identity established above is sometimes rearranged as

\[
E \left[ X^2 \right] = \mu_X^2 + \sigma_X^2.
\]

• In Normal \(N(\mu, \sigma^2)\) samples 95% of the values lie within 1.96\(\sigma\) of \(\mu\). Thus the variance of an estimator (which is typically at least approximately normally distributed) is inversely related to the accuracy - an estimator with a small variance will, with high probability, be close to its mean. If this mean is in fact the quantity one is trying to estimate, we say the estimator is unbiased.
• Example: Let \( Y_1, \ldots, Y_n \) be a sample of mortar measurements (independent, and all randomly drawn from the same population). If the population mean is \( \mu_Y \) and the variance is \( \sigma^2_Y \), then the mean and variance of the sample average

\[
\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i
\]

are (why?)

\[
\mu_{\bar{Y}} = \frac{1}{n} \sum_{i=1}^{n} E[Y_i] = \mu_Y,
\]

\[
\sigma^2_{\bar{Y}} = \sum_{i=1}^{n} VAR\left[\frac{Y_i}{n}\right] \quad \text{WHY?} \quad \sum_{i=1}^{n} \frac{VAR[Y_i]}{n^2} = \frac{\sigma^2_Y}{n}.
\]

Thus \( \bar{Y} \) is an unbiased estimator of \( \mu_Y \); it can be shown that in normal samples, no other unbiased estimator has a smaller variance. But note the lack of robustness.
• Similarly (see derivation in text),

\[
S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2
\]

is an unbiased estimator of \( \sigma_Y^2 \). The \( n - 1 \) is the ‘degrees of freedom’; this refers to the fact that only the \( n - 1 \) differences \( Y_1 - \bar{Y}, \ldots, Y_{n-1} - \bar{Y} \) can vary, since \( \sum_{i=1}^{n} (Y_i - \bar{Y}) = 0 \) implies that

\[
Y_n - \bar{Y} = - \sum_{i=1}^{n-1} (Y_i - \bar{Y}).
\]

• We will very often have to estimate the variance of other estimates, typically the effects of certain ‘treatments’, or other factors, on experimental units (e.g., effect of the modification on the mortar). In all cases the estimate will take the form

\[
\hat{\sigma}^2 = \frac{SS}{df},
\]

where \( SS \) is a sum of squares of departures from the estimate, and \( df \) is the degrees of freedom.
In many such cases we will be able (in theory, at least) to represent $SS$ as

$$SS = \sum_{i=1}^{df} Z_i^2,$$

where $Z_1, ..., Z_{df}$ are independent $N(0, \sigma^2)$ r.v.s.

This defines a chi-square r.v.:

$$\frac{SS}{\sigma^2} = \sum_{i=1}^{df} \left( \frac{Z_i}{\sigma} \right)^2 = a \text{ sum of } df \text{ squares of independent } N(0, 1)'s \sim \chi^2_{df}.$$

See some densities (Figure 2-6 of the text).

- Note $E \left[ \chi^2_{df} \right] = df$ (why?); also $VAR \left[ \chi^2_{df} \right] = 2df$. Thus

$$\hat{\sigma}^2 \sim \sigma^2 \frac{\chi^2_{df}}{df}$$

with $E \left[ \hat{\sigma}^2 \right] = \sigma^2$. What is the variance of this estimator? How does it vary with $df$?
3. Comparing two means

- Testing whether or not certain treatments have the same effects often involves estimating the variance among the group averages, and comparing this to an estimate of the underlying, ‘residual’ variation. The former is attributed to differences between the treatments, the latter to natural variation that one cannot control. Large values of the ratio of these variances indicates true treatment differences. Mathematically, one computes

\[ F_0 = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}, \]

where \( \hat{\sigma}_j^2 \sim \sigma^2 \frac{\chi^2_{df_j}}{df_j} \) for \( j = 1, 2 \), and \( \hat{\sigma}_1^2, \hat{\sigma}_2^2 \) are independent r.v.s. Then \( F_0 \) is the numerical value of

\[ F \sim \frac{\chi^2_{df_1}/df_1}{\chi^2_{df_2}/df_2}, \]

where the two \( \chi^2 \)'s are independent; this is the definition of an \( F \) r.v. on \( (df_1, df_2) \) degrees of freedom: \( F \sim F_{df_1}^{df_2} \).
• Inferences involving just one mean (or the difference between two means) can be based on the $t$-distribution. One can generally reduce the problem to the following. We observe $Y_1, \ldots, Y_n \overset{ind}{\sim} N(\mu, \sigma^2)$. Then

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$ independent of $S^2 \sim \sigma^2 \frac{\chi^2_{n-1}}{n-1}$.

Here we use:

1. R.v.s $Y_1, \ldots, Y_n$ are jointly normally distributed if and only if every linear combination of them is also normal; this is the single most important property of the Normal distribution.

2. In Normal samples the estimate of the mean and the estimate of the variation around that mean are independent.
• Thus \( t_0 = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \) is the numerical value of the ratio of independent r.v.s

\[
t = \frac{\bar{Y} - \mu}{S/\sigma} \sim \frac{Z}{\sqrt{\chi^2_{n-1}/(n - 1)}} ,
\]
where \( Z \sim N(0, 1) \);

this is the definition of “Student’s t” on \( n - 1 \) d.f.

• You should now be able to show that \( t^2 \) follows an \( F_{1, n-1} \) distribution.

• Return now to the mortar comparison problem. Let \( \{Y_{ij}\}_{j=1}^{n_i} \) be the samples (\( i = 1 \) for modified, \( i = 2 \) for unmodified; \( n_1 = n_2 = 10 \)). Assume:

\[
Y_{11}, \ldots, Y_{1,n_1} \overset{\text{ind}}{\sim} N\left(\mu_{\text{mod}}, \sigma^2_1\right),
\]

\[
Y_{21}, \ldots, Y_{2,n_2} \overset{\text{ind}}{\sim} N\left(\mu_{\text{unmod}}, \sigma^2_2\right).
\]

First assume as well that \( \sigma^2_1 = \sigma^2_2 = \sigma^2 \), say. Then

\[
\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_{\text{mod}} - \mu_{\text{unmod}}, \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right).
\]
The ‘pooled’ estimate of $\sigma^2$ is

$$S_p^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}$$

$$\sim \sigma^2 \frac{\chi^2_{n_1-1} + \chi^2_{n_2-1}}{n_1 + n_2 - 2}$$

$$\sim \sigma^2 \frac{\chi^2_{n_1+n_2-2}}{n_1 + n_2 - 2}$$

(the sum of independent $\chi^2$’s is again $\chi^2$; the d.f. are additive). Thus

$$t = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_{mod} - \mu_{unmod})}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_{mod} - \mu_{unmod})}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} / S_p}$$

$$\sim t_{n_1+n_2-2}.$$

To test the hypotheses

$$H_0 : \mu_{mod} = \mu_{unmod}$$

$$H_1 : \mu_{mod} \neq \mu_{unmod}$$

we assume that the null hypothesis is true and
compute
\[
t_0 = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.
\]

If indeed \( H_0 \) is true we know that \( |t_0| \sim |t_{n_1+n_2-2}|. \) If \( H_0 \) is false then \( |t_0| \) behaves, on average, like a multiple of \( |\mu_{mod} - \mu_{unmod}| > 0. \) Thus values of \( t_0 \) which are large and positive, or large and negative, support the alternate hypothesis.

- The \textit{p-value} is the prob. of observing a \( |t_{n_1+n_2-2}| \) which is as large or larger than \( |t_0|: \)
  \[
p = P \left( t_{n_1+n_2-2} \geq |t_0| \text{ or } t_{n_1+n_2-2} \leq -|t_0| \right).
\]
  If \( p \) is sufficiently small, typically \( p < .05 \) or \( p < .01 \), we “reject \( H_0 \) in favour of \( H_1 \).”
> t.test(mod, unmod, var.equal=T)

Two Sample t-test

data: mod and unmod

t = -9.1094, df = 18, p-value = 3.678e-08

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
-1.4250734  -0.8909266

sample estimates:
mean of x  mean of y
16.764     17.922
• If $\sigma_1^2 = \sigma_2^2$ is not a reasonable assumption then

$$\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_{mod} - \mu_{unmod}, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right),$$

and $\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$ is an unbiased estimate of the variance, independent of $\bar{Y}_1 - \bar{Y}_2$. However, this variance estimate is no longer $\sim$ as a multiple of a $\chi^2$, so that

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

is not exactly a $t$ r.v. It is however well approximated by the $t_{\nu}$ distribution, (“Welch’s approximation”) with random d.f.

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}/n_1-1\right) + \left(\frac{S_2^2}{n_2}/n_2-1\right)}.$$
> t.test(mod, unmod)  # Welch's two-sample t-test

Welch Two Sample t-test
data: mod and unmod
t = -9.1094, df = 17.025, p-value = 5.894e-08
alternative hypothesis: true difference
in means is not equal to 0
95 percent confidence interval:
  -1.4261741 -0.8898259
sample estimates:
mean of x  mean of y
  16.764    17.922
4. Randomized block designs

- **Confidence intervals.** The following treatment seems general enough to cover the cases of practical interest. Any $t$-ratio can be written in the form

$$t = \frac{\hat{q} - q}{se(\hat{q})},$$

where $q$ is a quantity to be estimated (e.g. $\mu_1 - \mu_2$), $\hat{q}$ is an estimate of $q$ (e.g. $\bar{y}_1 - \bar{y}_2$) and $se(\hat{q})$ is an estimate of the standard deviation of $\hat{q}$ (e.g. $S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$). Then if $\pm t_{\alpha/2}$ are the points on the horizontal axis which each have $\alpha/2$ of the probability of the $t$-distribution lying beyond them, we have

$$1 - \alpha = P \left( -t_{\alpha/2} \leq t \leq t_{\alpha/2} \right)$$

$$= P \left( -t_{\alpha/2} \leq \frac{\hat{q} - q}{se(\hat{q})} \leq t_{\alpha/2} \right)$$

$$= P \left( \hat{q} - t_{\alpha/2} \cdot se(\hat{q}) \leq q \leq \hat{q} + t_{\alpha/2} \cdot se(\hat{q}) \right),$$
after a rearrangement. Thus, before we take the sample, we know that the random interval

\[ CI = [\hat{q} - t_{\alpha/2} \cdot se(\hat{q}), \hat{q} + t_{\alpha/2} \cdot se(\hat{q})] \]

will, with probability \(1 - \alpha\), contain the true value of \(q\). After the sample is taken, and \(\hat{q}, se(\hat{q})\) calculated numerically, we call \(CI\) a 100 \((1 - \alpha)\) % confidence interval.

- In the mortar example, the difference in averages was \(\hat{q} = 16.764 - 17.922 = -1.158\); then from the computer output,

\[-9.1094 = t = \frac{-1.158}{se(\hat{q})} \Rightarrow se(\hat{q}) = \frac{-1.158}{9.1094} = .1271.\]

There were 18 d.f., and so for a 95% interval we find \(t_{18, \alpha/2}\) on \(R\):

\[ q t(.975, 18) \]

\[ [1] 2.100922. \]

Thus \(CI = -1.158 \pm 2.1009 \cdot .1271 = -1.158 \pm .2670 = [-1.4250, -.8910]\), in agreement with the computer output.
Sample size calculations. The power of a test is the probability of rejecting the null hypothesis when it is false. Suppose that we would like a power of at least .99 when the mortar means differ by $\delta = \mu_{mod} - \mu_{unmod} = .5$. Furthermore, suppose that we will reject $H_0$ if the p-value is $< .05$. (We ‘use $\alpha = .05$’. ) We need an estimate of $\sigma$, here I’ll use $\sigma = .4$, which is somewhat larger than $S_p$ was.

```r
> power.t.test(delta = .5, sd = .4, 
sig.level = .05, power = .99, type = "two.sample", 
alternative = "two.sided")
Two-sample t test power calculation
n = 24.52528
delta = 0.5
sd = 0.4
sig.level = 0.05
power = 0.99
alternative = two.sided
NOTE: n is number in *each* group
```

Thus in a future study, if $\sigma = .4$ is accurate, we will need two equal samples of at least 25 each in order to get the required power. Use > help(power.t.test) to get more details on this function.
• **Paired comparisons.** Suppose that the mortar data had arisen in the following way. Ten samples of raw material were taken, and each was split into two. One (randomly chosen) of the two became modified mortar, the other unmodified. This is then a ‘randomized block’ design - the raw materials are blocks, and observations within a block are expected to be more homogeneous (alike) than those in different blocks. Since the blocks (raw materials) might well affect the strength of the mortar, we should compare treatments using the same raw material, i.e. we put

\[ d_j = Y_{1j} - Y_{2j} \sim N(\mu_d = \mu_{mod} - \mu_{unmod}, \sigma_d^2) \]

and make inferences about \( \mu_d \) by carrying out a one-sample or paired \( t \)-test.
> t.test(mod, unmod, paired=TRUE)
  Paired t-test
data: mod and unmod
t = -10.2311, df = 9, p-value = 2.958e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -1.4140405  -0.9019595
sample estimates:
mean of the differences
  -1.158
• Testing the normality assumption. If \( X \sim N(\mu, \sigma^2) \), then 
\[
Z = \frac{X - \mu}{\sigma} \sim N(0, 1)
\]
and (with \( \Phi(x) = P(Z \leq x) \)):
\[
P \left( X \leq \sigma \Phi^{-1}(i/n) + \mu \right) = P \left( Z \leq \Phi^{-1}(i/n) \right)
= \Phi(\Phi^{-1}(i/n))
= i/n.
\]

On the other hand, if \( x(1) < x(2) < \cdots < x(n) \) are the order statistics of a sample from a population of \( X \)'s, then \( i/n \) is the sample-based estimate of \( P \left( X \leq x(i) \right) \). Thus, if the population is indeed Normal, we expect
\[
x(i) \approx \sigma \Phi^{-1}(i/n) + \mu,
\]
so that a plot of \( x(i) \) against the Normal quantiles \( \Phi^{-1}(i/n) \) should be approximately a straight line, with slope \( \sigma \) and intercept \( \mu \). In practice \( i/n \) is replaced by something like \( (i - .5)/n \) to avoid \( \Phi^{-1}(i/n) = \infty \) when \( i = n \).
> diff = mod-unmod
> qqnorm(diff)
> qqline(diff)
> cor(sort(diff),qnorm((.5:9.5)/10))
[1] 0.9683504

Some packages (e.g. MINITAB) will carry out a formal test, rejecting the null hypothesis of Normality if the correlation between \( x(i), \Phi^{-1}((i - .5)/n) \) is too low.