Basic Data Analysis

Stephen Turnbull

Business Administration and Public Policy

Lecture 8: May 31, 2013

Abstract

We continue interval estimation and hypothesis testing.

We introduce correlation and regression.

Interval estimates

- In opinion polls, you will often see estimates qualified with an estimate of the likely deviation from the truth, such as "$45\% \pm 3\%$ of the voters plan to vote for the LDP."
- This is called an *interval estimate* (区間推定) or *confidence interval* (信頼区間). It is interpreted as $0.42 \leq \alpha \leq 0.48$ (α is the fraction of LDP voters).
- Where does the $\pm 3\%$ come from? Can we *guarantee* that α is truly in that range? No.
- We are confident that it is, and can quantify our confidence in probability-like terms, such as a *90% confidence interval*.

Confidence is *not* probability

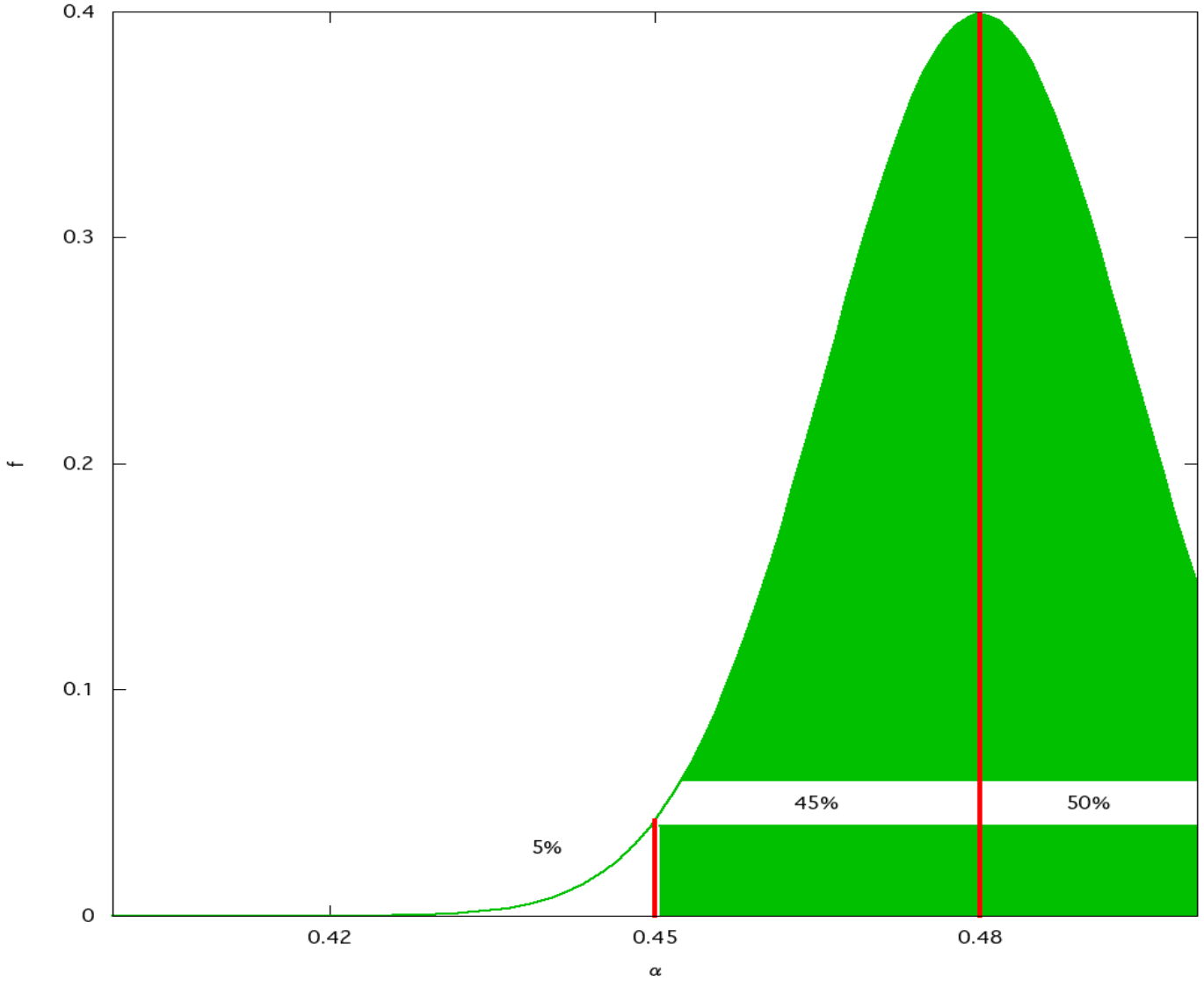
- We quantify "confidence" in probability-*like* terms.
- However, it is *not* a probability. If we estimate the mean by $\bar{X} \pm .03$, the true μ either *is* in the range, or it *is not*. We don't know which is true, but it's *not* random!
- One way to think about it is to try to compute a probability. Suppose our distribution is normal. Then to compute a probability we need to know the mean. But our confidence interval says that the mean is somewhere between 1.5 and 3.2. What does

$$\int_{-\infty}^{2} \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{z-(\text{somewhere between 1.4 and 3.2})}{2}\right)^2} dz$$

mean?

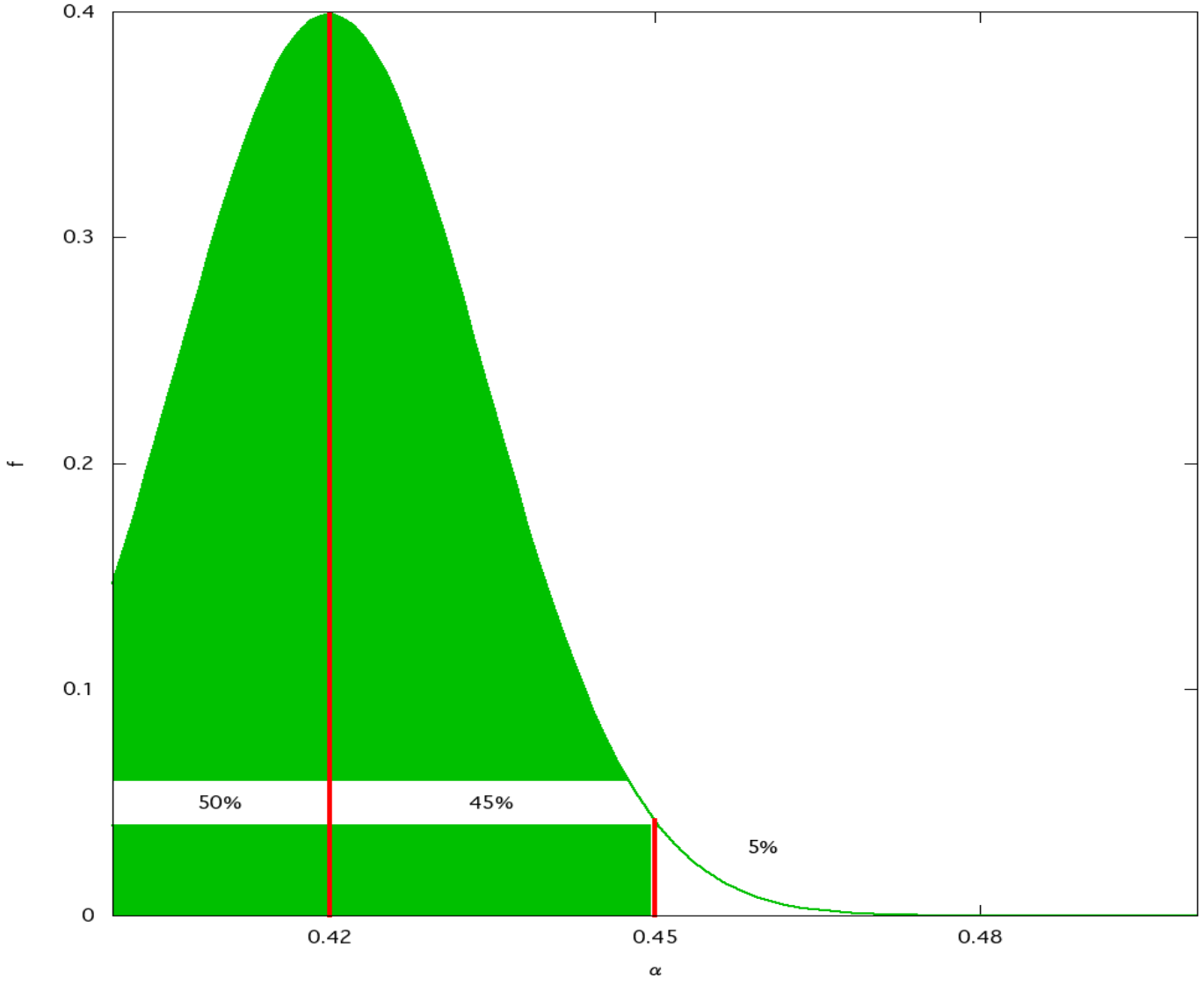
Computing confidence: upper bound

We are 95% confident that α is smaller than 0.48 because if α were 0.48, the probability of $\hat{\alpha}$ being 0.45 or more is 0.95. It is *unlikely* that we observe $\hat{\alpha}$ as small as 0.45, *given* the estimated mean $\hat{\alpha}$.



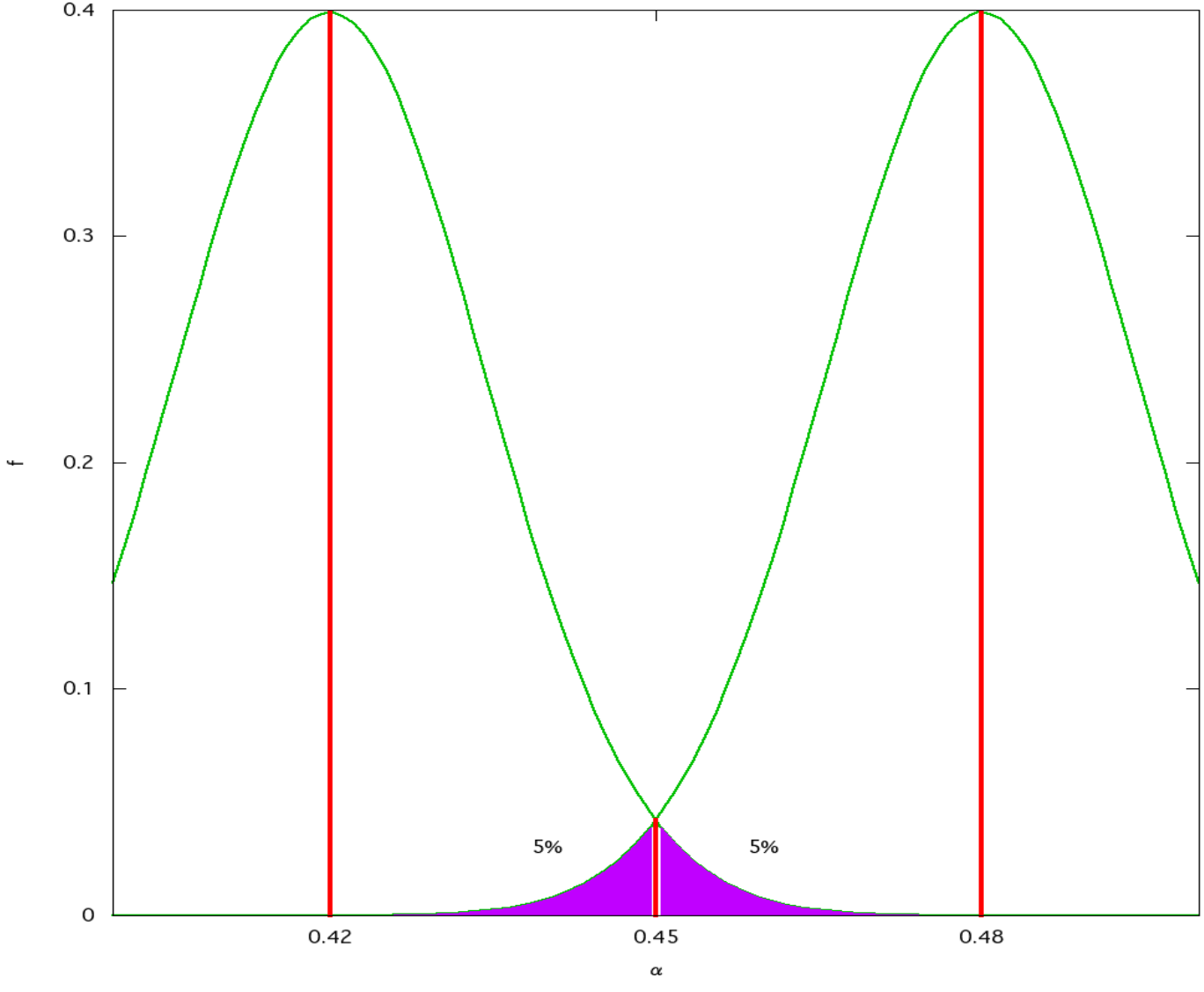
Computing confidence: lower bound

We are 95% confident that α is larger than 0.42 because if α were 0.42, the probability of $\hat{\alpha}$ being 0.45 or less is 0.95.



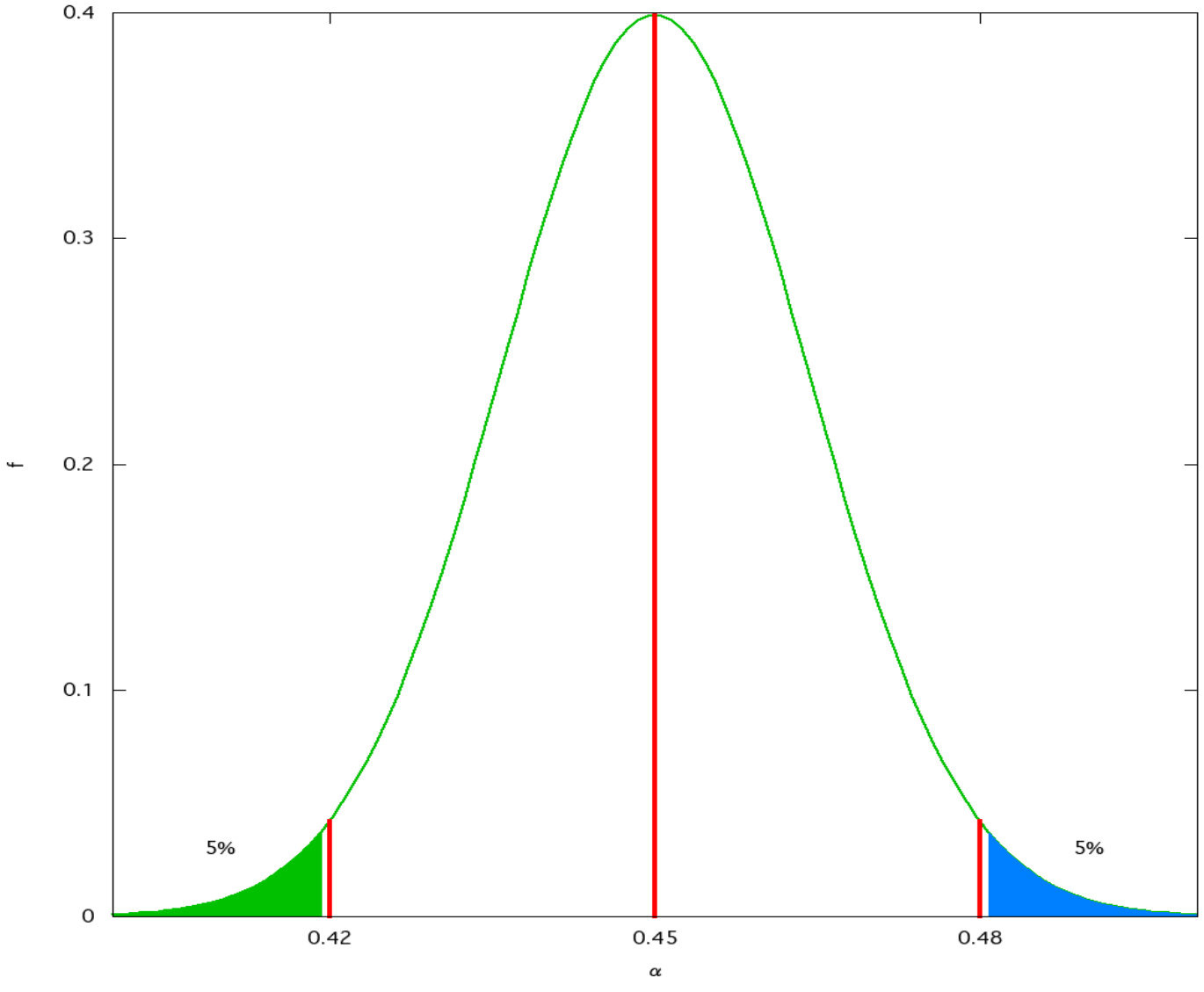
A symmetric interval

We are 90% confident that α is larger than 0.42 but lower than 0.48. The deviation probabilities ("probability of deviation outside the limit") are equal.



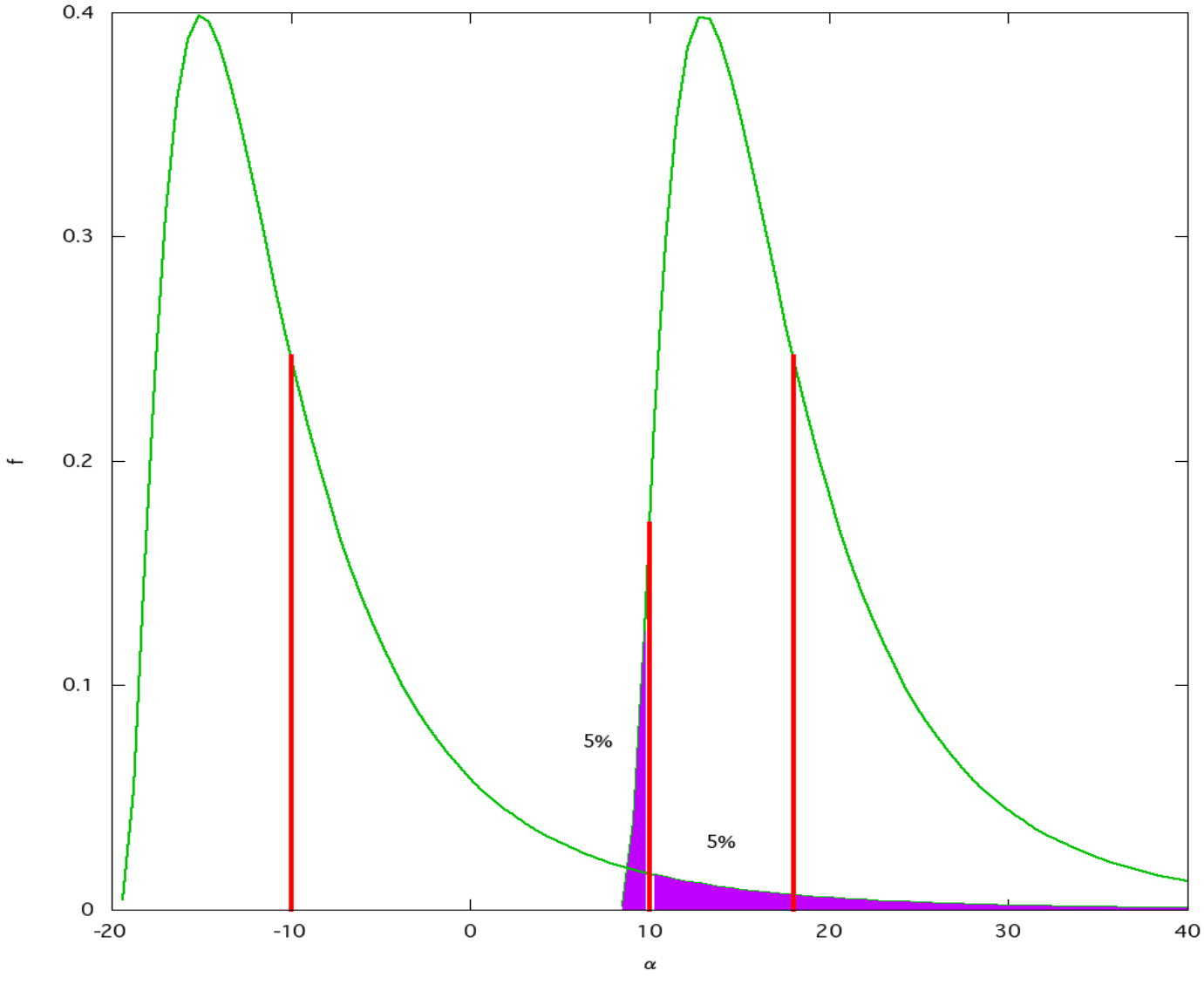
How not to compute a confidence interval

This is the wrong way to compute a 90% confidence interval; it assumes that $\alpha = 0.45$, *i.e.*, $\hat{\alpha}$ is known to be correct. But α is unknown.



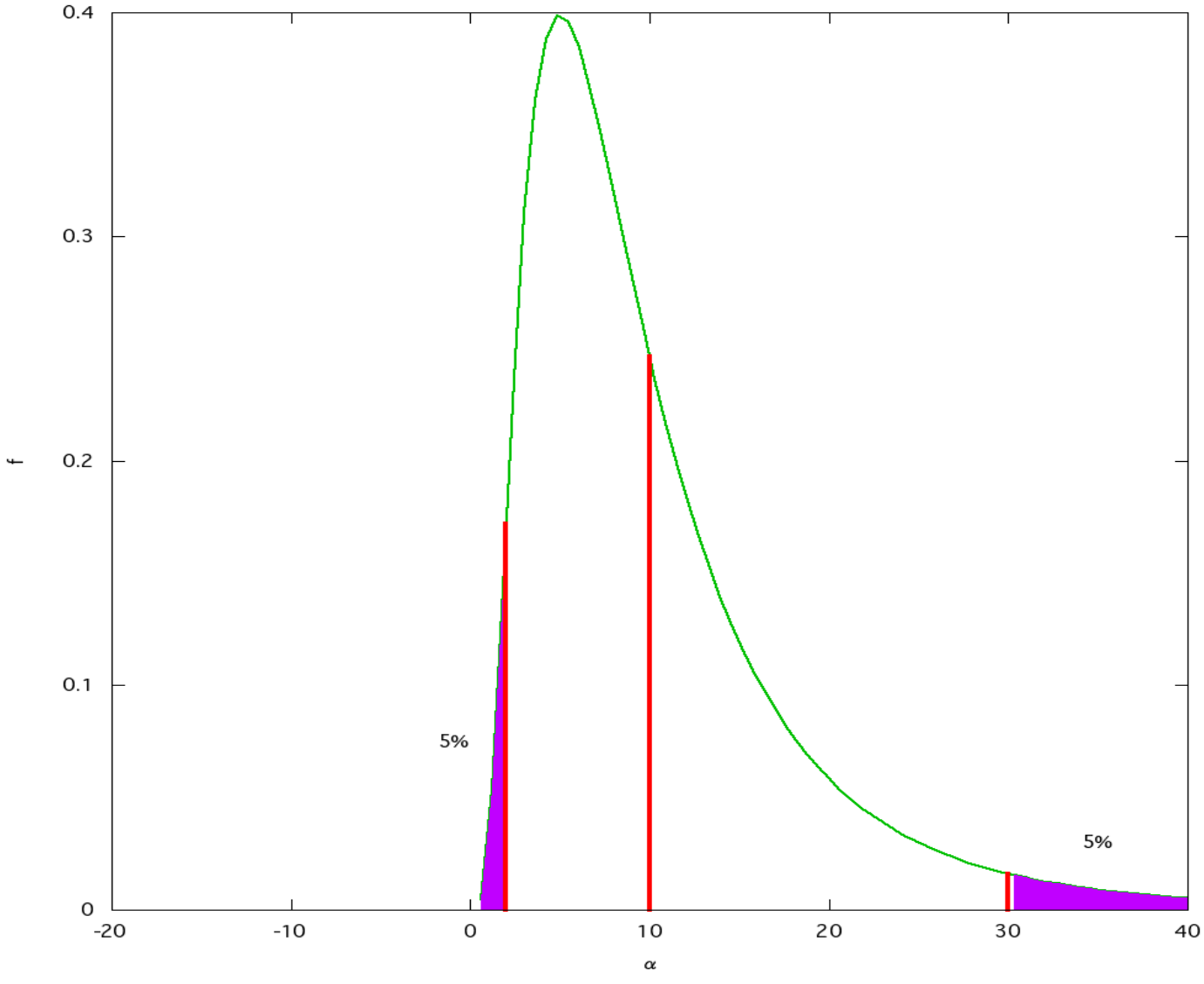
A skewed distribution

We call this an *asymmetric confidence interval* because the deviation probabilities are equal, not the distance from the mean. It's the right way to do it.



Incorrect interval for skewed distribution

Note the distances to the upper and lower bounds are reversed.



Statistical test of theory

- A statistical test ((統計的)検定) requires two things.
- A *quantitative model* of the theory (often called "the domain model").
- A *statistical model* of dispersion in the data.

Statistical models

- The *domain model* is expressed as an equation (or several equations).
- For example, a model of costs sufficient for testing returns to scale might be

$$C_t = a + bQ_t + cQ_t^2,$$

  where C_t is the expenditure in period t and Q_t is the quantity produced in period t, and a, b, and c are model parameters.
- Uppercase Latin letters denote data, and lowercase Latin letters are model parameters.

Statistical models: examples

- The *statistical model* makes the equality uncertain. It involves introducing a random variable in the domain model.
- The *linear regression model* is simplest, just add randomness:

$$C_t = a + bQ_t + cQ_t^2 + \epsilon_t.$$

- The *measurement error model* assumes the data is measured inaccurately:

$$C_t + \epsilon_{Ct} = a + b(Q_t + \epsilon_{Qt}) + c(Q_t + \epsilon_{Qt}))^2.$$

- The *random coefficients model* assumes the parameters are random! Like this:

$$C_t = (a + \eta_a) + (b + \eta_b)Q_t + (c + \eta_c)Q_t^2.$$

- The Greek letters ϵ and η denote unobserved random variables ("errors").

Verifying theory

- In many important applications we have a "theory" we want to confirm (or disprove):
  - There is no gender discrimination in an certain organization.
  - English ability is valued by companies.
  - A firm's production shows decreasing returns to scale.
- To work with these statistically we must have a *quantitative model* of the theory.

Quantifying the hypothesis

- We need to *measure* something, and *compare it to another value*. This is the *hypothesis* ((統計的)仮説).
  - No gross gender discrimination in labor markets: We measure the "attitude" toward each gender by the average wage, W_i, $i = m, f$ (average wage of group i, m is male, f is female). No discrimination means $W_m = W_f$. (What do we mean by "*the* wage"?
  - English ability is valued by companies: Measure "value" by wage. The hypothesis is $W_1 > W_0$, where W_1 is wage of an employee with a qualification, W_0 the wage without.
  - A firm's productivity can be measured as the (negative of) the cost function $C(q) = a + bq + cq^2$. It shows decreasing returns to scale when $c > 0$.

Modeling voting

- The quantitative model is simple: we look at the fraction of people who say "yes" to the question. Each either says "yes" ($X_i = 1$) or "no" ($X_i = 0$), and the fraction then is the "average" vote: $\alpha = \frac{1}{n} \sum_{i=1}^{n} X_i$.
- The statistical model is based on *random sampling*. That is the reason for variation is not that "people change their minds," but rather that "whether a person is asked or not is random".
  - This *almost never* gives a *perfectly* representative sample.
  - On *average* it gives a fairly representative sample.

Modeling production

- The quantitative model is *economic profit maximization*, which implies *cost minimization* and the existence of a *cost function* (*i.e.*, a map not from inputs and their prices to expenditure, but a map from *output* and input prices to expenditure).
- A simple statistical model is *weather damage to crops*; every year there is some, but it varies.
- The important point is that *weather damage depends on random weather, not on our inputs*. Then $C(q) = \bar{C}(q) + \alpha$, where $\alpha > 0$ is the random weather damage.
  - Then $\alpha = C(q) - \bar{C}(q)$, and if $\alpha \sim N(\mu, \sigma)$, then deviations from projected cost (*i.e.*, before adjusting for weather damage) are distributed $N(\mu, \sigma)$!

Testing hypotheses

- What does it mean to test a hypothesis? (仮説検定)
- First we need a statistical model, as explained. Let's consider the voting model, and suppose the question was "will you vote LDP in the next election?" To make it interesting (and simple), assume a "no" answer implies voting for the DPJ.
- Let's consider *two* simple hypotheses.
  1. The parties have the same support in the population of voters.
  2. The DPJ is winning.
- They seem closely related, but there is a very important technical difference. This difference is based on the fact that Hypothesis 1 is *symmetric* in the two parties, while Hypothesis 2 is actually *asymmetric* (from a certain point of view).

The null hypothesis H_0

- Note that we numbered our hypotheses. This is common and useful practice in applying statistics to practical problems. But be careful not to become confused, because there are two "special" numbered hypotheses, the "null hypothesis" H_0 (帰無仮説), and the "alternative hypothesis" H_1 (対立仮説).
  - H_1 is a *different* usage from Hypothesis 1 above.
- The "0" in H_0 is like the 0 of a graph: it is the *origin*, the point of reference.
- Specifically, the *null hypothesis* is *the quantitative expression of a hypothesis as the specific value of a parameter of the statistical model used to compute probabilities of observable events.*
- The observable events are expressed relative to the data set.

What are our null hypotheses?

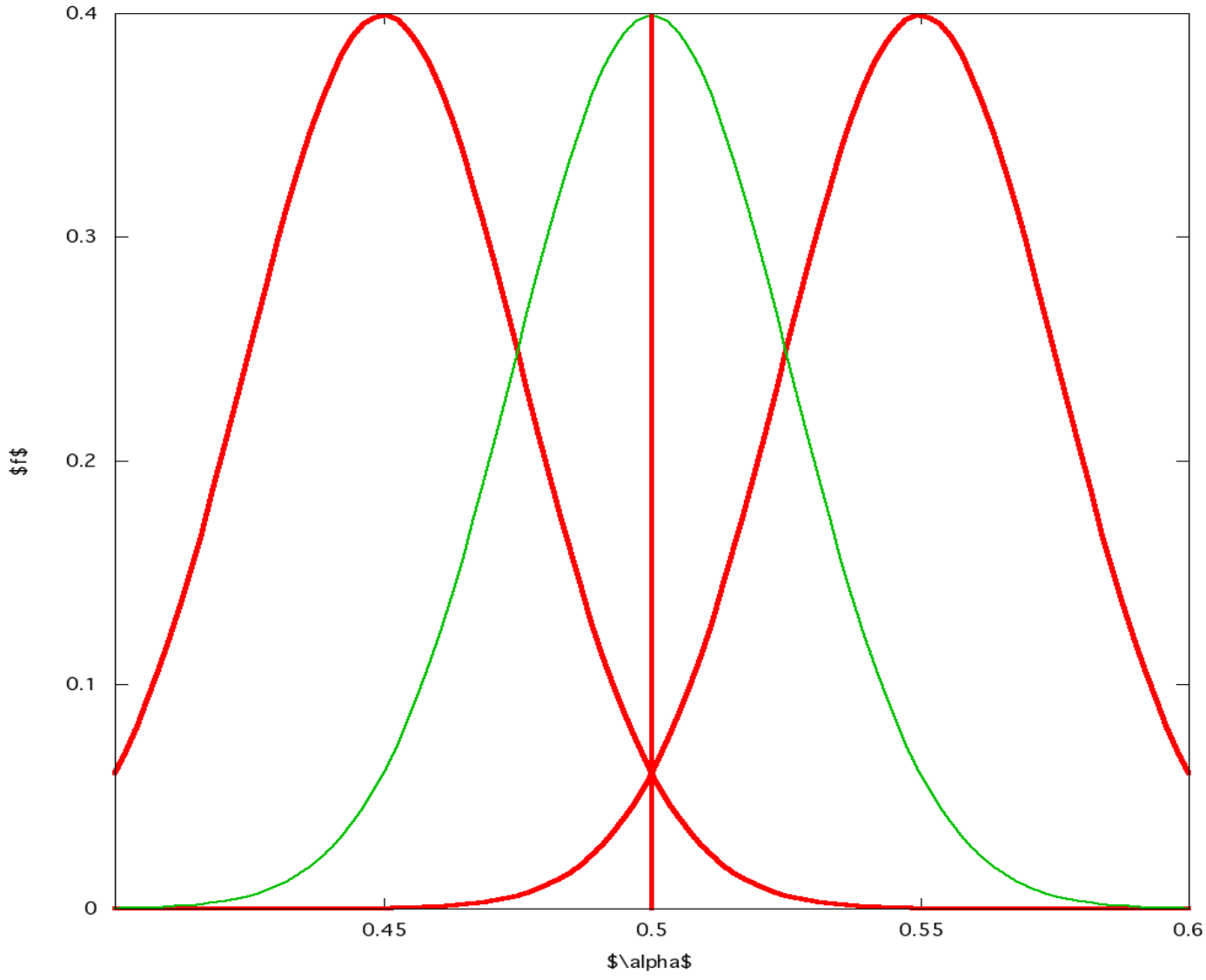
- In Hypothesis 1, "the parties have the same support in the population of voters," the null hypothesis should be obvious: $H_0 : \alpha = 0.5$.
  - The alternative hypothesis is $H_1 : \alpha \neq 0.5$.
  - Note that H_1 is *almost always* satisfied by the data.
  - But *it cannot be used to compute probability statements about the data.*
- Both $\alpha < 0.5$ and $\alpha > 0.5$ satisfy H_1 (it's *two-sided*).
- In Hypothesis 2, "the DPJ is winning," it is not obvious how to get a probability statement! *There is no obvious specific value of α to use.*
  - This is related to Hypothesis 2 being *one-sided.*

One-sided tests

- We can't use $\alpha < 0.5$, because we can't compute with it.
- Picking $\alpha = 0.45$ is not helpful for two reasons.
  - Technically speaking, since it's the maximum likelihood, it can never be rejected.
  - Since it's necessarily inaccurate, it has no theoretical claim on our attention.
- The way out: make "the DPJ is winning" the *alternative hypothesis*.
  - This fits with the ambiguity.
- What is H_0? We can't calculate without it!

H_0 for one-sided tests

$H_0 : \alpha = 0.5$ is the null hypothesis to use. It gives the highest probability of the observed data among null hypotheses that mean "the DPJ is *not* winning."



Conducting the test

- The basic result of a test is *pass* or *fail*. In statistics, it is to *accept the null hypothesis* (採択 - implying the alternative is rejected) or to *reject the null hypothesis* (棄却 - ききゃく - and the alternative is accepted).
- The procedure is to pick a *significance level* (水準) or *critical P-value*, such as 0.05 (5%).
- Based on the parameter value(s) in the *null* hypothesis, compute a *critical region* E (棄却域 - an event) such that $P[\bar{X} \in E] = 0.05$. The critical region may be defined by
  - an *upper critical value*, and anything greater rejects H_0,
  - a *lower critical value*, and anything less rejects H_0, or
  - both, and anything *outside* those bounds rejects H_0, or
  - some more complicated set (but we don't deal with that!)

Testing the election

- Our theory: the DPJ is preferred by the voters.
- Define α to be the fraction that prefer the LDP. Then $H_0 : \alpha = 0.5$ and $H_1 : \alpha < 0.5$. The theory corresponds to H_1.
- Statistical model: $\bar{X} \sim N(\alpha, 0.0182)$ (same σ as before).
- Let the significance level be 0.05.
- The lower critical value $\underline{\alpha}$ satisfies $P[\bar{X} \leq \underline{\alpha}] = 0.05$.
- Standardizing, $0.05 = P[z \leq \frac{\underline{\alpha}-0.5}{0.0182}]$ where $z = \frac{\bar{X}-0.5}{0.0182}$.
- The critical value of z is -1.65, so $-1.65 = \frac{\underline{\alpha}-0.5}{0.0182}$ and $\underline{\alpha} = 0.5 - (1.65)(0.0182) = 0.47$.
- Since the observed value is $0.45 < 0.47$, we *reject* H_0 and accept H_1, and conclude that the DPJ is winning.

Hypothesis testing and interval estimation

- The similarity of computation is no accident.
- Any hypothesis test can be seen as constructing a confidence interval.
- We didn't discuss one-sided confidence intervals, but they are sometimes useful. *E.g.*, consider if you are working for the LDP and want to estimate the probability of winning: "95% confident we win."

Type I and Type II errors

- Because of sampling and other random factors, hypothesis tests are not 100% reliable. Although in most cases we can never verify the truth, conceptually we can classify in this way:

Null hypothesis H_0 is

	True	False
Accepted	OK	Type I error
Rejected	Type II error	OK

Table 1: Hypothesis testing errors

- Note the distinction between *accept* and *true*, and similarly *reject vs. false*. Unfortunately researchers often say "true" when they mean "accepted"—be careful!

Significance and power of tests

- P[Type I error] is called the *power* of the test, often denoted by β. Low β is good.
- P[Type II error] is called the *significance* of the test, often denoted by α. Low α is good.
- Making α smaller will increase β and vice versa. (You can choose either one, and the underlying distribution then determines the other.)
- Making N bigger allows you to decrease both α and β (or more likely, keep α the same and decrease β).

Example: A Complex Hypothesis Test

- Consider a simple example of budgeting a political campaign. You are the campaign director for a candidate for the Diet.
- Your candidate is a strong candidate usually, but this year she faces a tough race because her opponent is a charismatic former "idol" and the voters are mad at her party.
  - You worry that she might lose, but ...
  - You would like to save money for her *next* campaign, or to share with weaker candidates she favors.
- You have **4 weeks** until the election, and the results of two polls taken last week and this week.

The Poll Results

- Each poll surveyed 400 likely voters, and you believe they are well-designed surveys of randomly-selected likely voters.
  - This means you may assume that the voters are independent and identical random draws from the population of voters.
- In last week's poll, your candidate received 48% of the vote (and the other candidate 52%). The standard error of the estimate is 1.5%.
- In this week's poll, your candidate received 49% of the vote (and the other candidate 51%). The standard error of the estimate is 1%.

The Basic Strategy

- If the poll results are accurate and you continue spending at the same rate, your candidate should gain 1% per week. At the election in 4 weeks your candidate should win comfortably, 53% to 47%.
- After discussing with your client, you have decided that you don't need to spend more money this week if you are confident that she will get at least 52% of the vote based on current trend and statistical analysis.

What is the Hypothesis to Test?

- We'd like to have a model of how voters make up their minds, and apply our data to that model and the question of whether at election time more than 52% of the voters will vote for our candidate. This is difficult. Among other things, the data for *last* week could be considered to have an implication about the true fraction intending to vote for our candidate as of *this* week. Also, we know the estimate of 49% this week may be in error.
- So we will focus on the simple question of "is our candidate gaining votes fast enough to reach 52% by the election?"
  - We assume the estimate this week is accurate.
- Then we need the *gain from last week to this week* to be at least $\frac{3\%}{4} = 0.75$.

What is H_0?

- We need to formulate a *null hypothesis*.
- Let μ_1 be the true fraction of voters voting for our candidate last week, and μ_2 the fraction this week.
- We want to compare $\mu_2 - \mu_1$ to 0.75.
- Should our null hypothesis be
  1. $H_0 : \mu_2 - \mu_1 = 0.75$ (two-sided), or
  2. $H_0 : \mu_2 - \mu_1 \geq 0.75$ (one-sided), or
  3. $H_0 : \mu_2 - \mu_1 \leq 0.75$ (one-sided)?
- It's easy to reject the two-sided version; we care whether she wins or loses, not whether the election is a tie or decisive.

What Does $H_0 : \mu_2 - \mu_1 \geq 0.75$ Mean?

- Technically, if we *reject* the hypothesis, we are very confident that our candidate is going to have less than 52% at election time.
- We will *reject* only if $\mu_2 - \mu_1$ is *significantly less* than 0.75.
- We *may* accept the hypothesis if $\mu_2 - \mu_1$ is *only somewhat less* than 0.75. It doesn't need to be more!

What Does $H_0 : \mu_2 - \mu_1 \leq 0.75$ Mean?

- Technically, if we *accept* the hypothesis, we are very confident that our candidate is going to have more than 52% at election time.
- We will *reject* only if $\mu_2 - \mu_1$ is *significantly more* than 0.75.
- We *may* accept the hypothesis if $\mu_2 - \mu_1$ is *only somewhat more* than 0.75. It doesn't need to be less.

Data model

- We need to know how to calculate our standard error of the estimate.
- However, $\hat{\mu}_1$ and $\hat{\mu}_2$ are random variables with the same state space. They may be dependent.
- A model in which they are *independent* is one where μ_1 and μ_2 are fixed values, and there is only sampling error.
- This model is somewhat plausible, but we need to be careful, because we assume their values are *related by a trend*. The statistical model therefore could easily involve *relation by correlation*.

Comparing distributions for equality

- Recall the two datasets of grades for two instances of the same class, with the same class size:

  AM Class B C A A A B A C

  PM Class A B C B B C A B

  and distributions

	D	C	B	A
AM	0	2	2	4
PM	0	2	4	2

- *As distributions*, we concluded that the morning's was clearly *higher* than the afternoon's, because the CDF of the AM class was *everywhere to the right* of the CDF of the PM class.

- But what if we wanted to determine whether the distributions of abilities of the students were "really" different?

Comparing the whole distributions

- We want to develop a statistic that determines whether two distributions differ *significantly*. Here they can be different in any way: not just location, but dispersion, skewness, kurtosis, or even at some particular value, differences can be significant.
- We take as the null hypothesis that (1) AM is the standard of comparison, and (2) $f_{PM}(x) = f_{AM}(x)$ for all x in A, B, C, and D.

Developing a statistic

- Algebraicly, the distributions are "different at x" if $f_{PM}(x) - f_{AM}(x) \neq 0$. But if we use the algebraic difference, two differences may have the opposite sign and cancel out. We use the usual trick of measuring difference with the *square* of the algebraic difference.
- There is also a scaling problem: large values probably have more "natural variation." We rescale by dividing by the "standard" frequency.
- Adding up the variations gives the χ^2 statistic:

$$\chi^2 = \sum_{x \in \{D,C,B,A\}} \frac{(f_{PM}(x) - f_{AM}(x))^2}{f_{AM}(x)}.$$

What about the D cell?

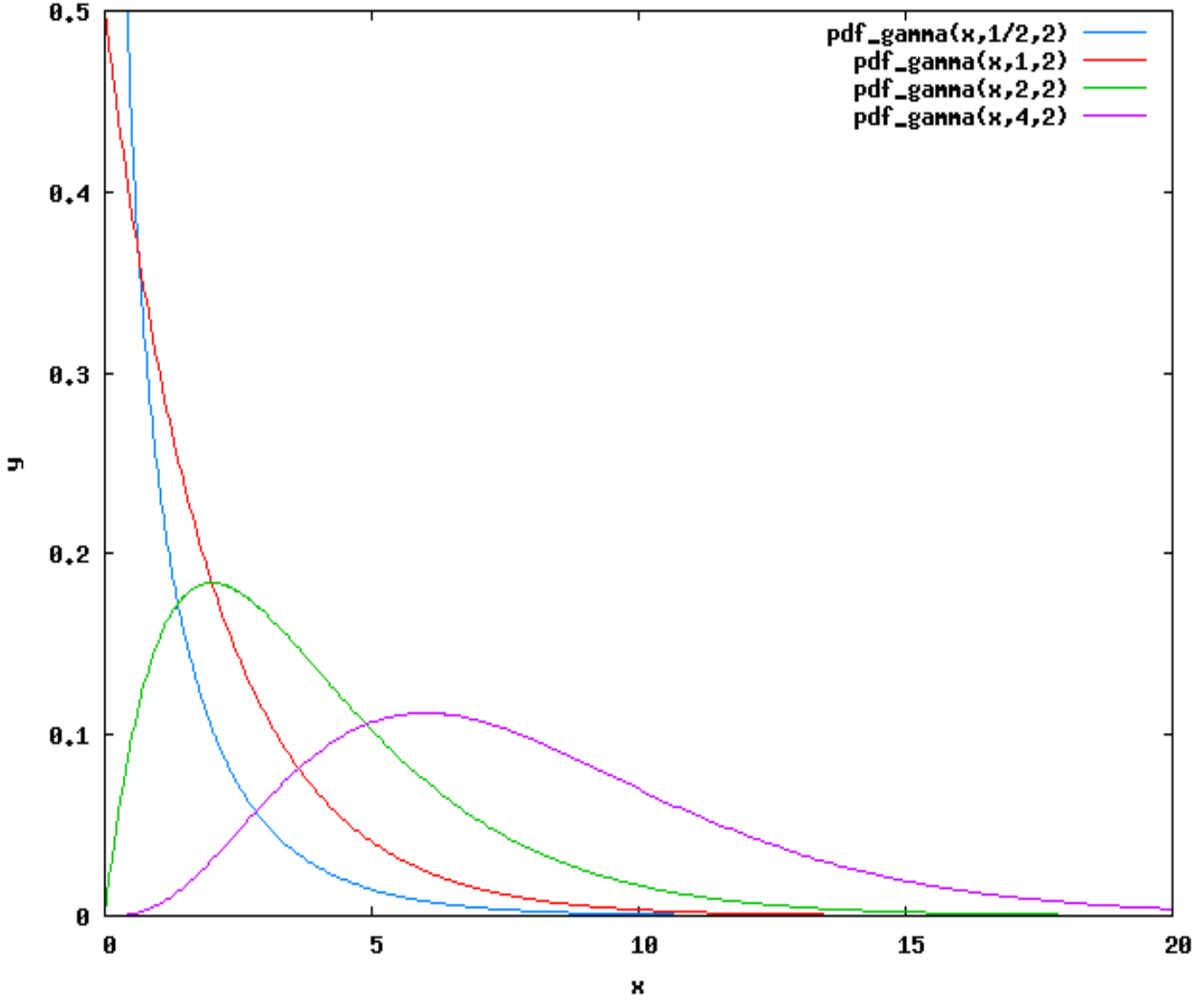
A typical problem of applying statistical theory to real data.

- The D cell has a frequency of 0 in the morning, so the formula needs to include $\frac{f_{PM}(D)-0}{0}$, which is normally mathematical nonsense. What to do?
- In this case, $f_{PM} = f_{AM}$, so this is 0/0, still nonsense. But note that for all $x \neq 0$, $\frac{x-x}{x}$ makes sense and is equal to 0. It is plausible here to assume continuity, and make the first term of the χ^2 sum $0^2 = 0$.
- In the case of one or the other being non-zero, we can choose that one as the standard of comparison, or the average.
- Note that in the case of one being non-zero, we must include it in the calculation and in the degrees of freedom. So we should do the same in the case of both zero.

The χ^2 distribution

- If the standardized deviations are independently and standard normal distributed, the χ^2 statistic for $n+1$ cells has a χ^2 distribution with n degrees of freedom, denoted χ^2_n.
- You can look up the table in the back of the book, or use functions in your statistical software (including Excel).
- But what does "n degrees of freedom" mean?
- Since each additional r.v. has positive mean $\mu = 1 = \mathcal{E}[z^2] = \mathcal{V}[z] + \mathcal{E}[z]^2$ and positive variance $\sigma^2 = 2 = \mathcal{E}[z^4] - \mathcal{E}[z^2]^2$ (recall that the kurtosis of a normal random variable is 3), so a χ^2_n r.v. has mean n and variance $2n$. However, for several reasons (including the skewness), the *distribution* of a χ^2_n cannot be computed by a simple transformation of a χ^2_1 r.v.

χ^2 densities, $n = 1, 2, 4, 8$



Degrees of freedom

- The simple answer is "a parameter for the χ^2 distribution" among others.
- When comparing to a fully specified distribution, it's basically the number of cells. In fact n = number of cells - 1.
- Why subtract 1? Because given the relative frequency of observations, if you know what fraction into each of n cells, you just subtract from 1 to find out how many are in the last cell. The last cell is constrained, and contributes no freedom. Each of the others contributes a degree of freedom.
- If you need to estimate parameters (*e.g.*, the mean or standard deviation of the specified distribution), you need to subtract an additional degree of freedom for each parameter estimated.

Are the classes different?

- We have 4 cells, so $n = 3$.
- The statistic is
  $\chi^2 = 0^2 + \frac{(2-2)^2}{2} + \frac{(4-2)^2}{2} + \frac{(2-4)^2}{4} = 0 + 1 + 2 + 1 = 4.$
- The P-value for a χ^2_3 variable at 4 is 0.23. If the AM distribution is the "true" distribution for the PM distribution, there is a 23% chance that the PM values (or ones even more different in the sense of χ^2) would appear after all.
- This chance is too high to be confident the difference is significant, and we accept the null hypothesis that the PM distribution is not different from the AM distribution,
  $H_0 : f_{PM}(x) = f_{AM}(x)$ for all cells x.

Continuous Distributions

- Suppose you have two samples from normal distributions (or any other distributions with large support). Then what?
- Just divide the support into cells, using the same cell ranges for both samples.
- Do the χ^2 test for those distributions.

Multivariate models

- In many important situations the variable of interest is not independent of everything else.
- *E.g.*, in our cost model both total cost and unit cost may depend on quantity.
  - A functional relationship.
  - Regression analysis is very helpful.
- In a population of people, although tall people are on average heavier than short ones, the relationship is not fixed and there are exceptions to a greater or lesser extent.
  - A statistical relationship.
  - Correlation analysis may be most revealing.

Covariance

- The *covariance* of two random variables is defined as a "mixed" central moment:

$$\mathrm{Cov}(X, Y) = \mathcal{E}[(X - \mathcal{E}[X])(Y - \mathcal{E}[Y]),$$

  often denoted σ_{XY}.

- This is well-defined for empirical moments as well as in probability theory.
- Like variance, it's not obvious what covariance really means (in terms of units).

Covariance matrix

- For more than 2 variables, it is useful to define the *covariance matrix*. For 3 r.v.s X, Y, Z

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{YX} & \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{ZX} & \sigma_{ZY} & \sigma_Z^2 \end{bmatrix}.$$

- The covariance matrix is symmetric ($\sigma_{ij} = \sigma_{ji}$).

Using the covariance matrix

- The covariance matrix is a building block in all multivariate analysis.
- For a linear combination of r.v.s $aX + bY$, we have $\mathcal{V}[aX + bY] = a^2\mathcal{V}[X] + 2ab\text{Cov}(X, Y) + b^2\mathcal{V}[Y]$.
- This generalizes: for $X = [X_1 \dots X_n]$ a sequence of random variables and a coefficient vector $a = [a_1 \dots a_n]$, we have $\mathcal{V}[a^T X] = a^T \Sigma a$, where Σ is the covariance matrix of the vector X.
- For practical interpretation, use the correlation coefficient.

Correlation

- The *correlation coefficient* of two random variables is a standardized version of the covariance:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

- $-1 \leq \rho_{XY} \leq 1$ for all r.v.s X and Y.
- If X and Y are independent r.v.s, then $\mathrm{Cov}(X, Y) = 0$, and it follows that $\rho_{XY} = 0$.
- The *correlation matrix* is constructed in the same way as the covariance matrix.
- A simple form of "data mining" is to collect observations on a large number of variables, construct the correlation matrix, and look for highly correlated (either positively or negatively) variables.

Finding data

- Two good places for getting various kinds of official economic and business data are
  - `http://www.e-stat.go.jp/`
  - `http://www.bea.gov/`

  Most countries now have similar sites. The U.S. Bureau of Economic Analysis's is probably the most complete and well-designed for easy access.
- Other good sources include central banks, the United Nations, the World Bank, the IMF, and the OECD.
- Industry-level and firm level data is harder to get. Besides the economic ministries of various countries, trade organizations for each industry may help.

Selecting data

- Most data-generating organizations present the data in various ways. For example, the U.S. BEA provides about 50 different tables describing national income and product accounts (NIPA) at various levels of detail (`http://www.bea.gov/national/nipaweb/SelectTable,asp?SelectedTable=N`, Section 1).
- Unless you are already quite expert in the particular data, it is a good idea to browse to see what presentations are available, and if there are any usage notes. (See example, next slide.)
- Factors that affect the usage of data for both business and economics usage include
  - inflation adjustment—with adjustment is called *real*, without is called *nominal* (most sources specify "real," but omit "nominal")

- seasonal adjustment (sometimes called *deseasonalized*)
- levels *vs.* indices (ratio of current level to a base level) *vs.* changes *vs.* growth rates (current ratio of change to level)
- **There are often different methods for adjustment; you should report *exactly* which method is used for the data you use**

Problematic data, example

- BEA NIPA Table 1.1.6 *Real Gross Domestic Product, Chained Dollars* provides this note:

  > Chained (2005) dollar series are calculated as the product of the chain-type quantity index and the 2005 current-dollar value of the corresponding series, divided by 100. Because the formula for the chain-type quantity indexes uses weights of more than one period, the corresponding chained-dollar estimates are usually not additive. The residual line is the difference between the first line and the sum of the most detailed lines.

  The data for year 1944 shows a 25% discrepancy!

- This means it is unsuitable for regression analysis involving multiple variables, since their values are inconsistent.

Retrieving and cleaning data

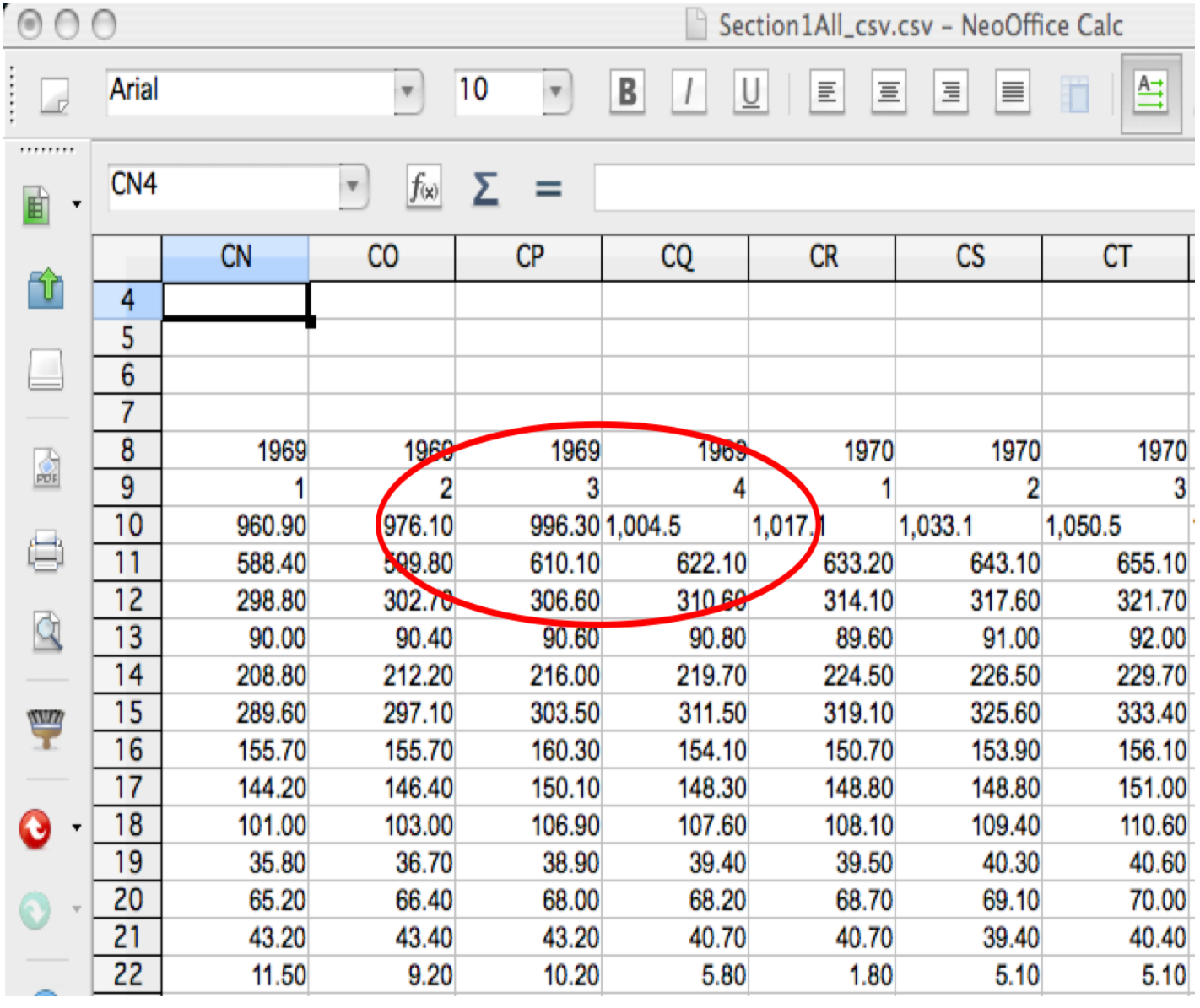
- Getting the data—for websites, just use Firefox, IE, *etc.*
- Statistical software is picky about format, even in spreadsheets
  - Each variable should be a *column*, not a *row*
  - There should be no extra titles, notes, *etc.*.
  - The only text should be variable names at the top of the column (see row `A` in `US-GDP-1947.1-2010.1.csv`). All other data should be *numerical*.

More about cleaning data

- Data sources often provide variables in *rows* (for large data sets, this makes browsing variables easier. Use the `TRANSPOSE` function in the spreadsheet to convert to columns.
- Remove extra titles like `A1:A6` in `Section1All_csv.csv`.
- The only text should be variable names at the top of the column. All other data should be *numerical*. Non-numerical data often appears as strange formatting in the spreadsheet (see next slide) and an error like `using type = "numeric" will be ignored` in R. Fixing this may be annoying, feel free to ask for help.
- Compare BEA NIPA Table 1.1.5 in `Section1All_csv.csv` with `US-GDP-1947.1-2010.1.csv` to see what needs to be changed.
- The spreadsheets are available on my homepage.

Sneaky problems

The problem with `lm` I had in class was due to text that snuck in through the spreadsheet. See how in cell `CQ10` the value is *left-justified* and *contains a comma*.



	CN	CO	CP	CQ	CR	CS	CT
4							
5							
6							
7							
8	1969	1969	1969	1969	1970	1970	1970
9	1	2	3	4	1	2	3
10	960.90	976.10	996.30	1,004.5	1,017.1	1,033.1	1,050.5
11	588.40	599.80	610.10	622.10	633.20	643.10	655.10
12	298.80	302.70	306.60	310.60	314.10	317.60	321.70
13	90.00	90.40	90.60	90.80	89.60	91.00	92.00
14	208.80	212.20	216.00	219.70	224.50	226.50	229.70
15	289.60	297.10	303.50	311.50	319.10	325.60	333.40
16	155.70	155.70	160.30	154.10	150.70	153.90	156.10
17	144.20	146.40	150.10	148.30	148.80	148.80	151.00
18	101.00	103.00	106.90	107.60	108.10	109.40	110.60
19	35.80	36.70	38.90	39.40	39.50	40.30	40.60
20	65.20	66.40	68.00	68.20	68.70	69.10	70.00
21	43.20	43.40	43.20	40.70	40.70	39.40	40.40
22	11.50	9.20	10.20	5.80	1.80	5.10	5.10

Statistical software: general

- Modern statistical software is generally designed to use an ASCII character set to encode statistical terms. Thus "χ^2" becomes something like "`chi2`", and "$\Phi(z)$" becomes "`normal_cdf(z)`".
- Linear equations are typically reduced to lists of data variables, with the computed coefficients labelled with the variable name instead of special symbols.
- Variables generally have multiletter (and number) names, rather than being a single character as is typical in algebra.
- The biggest hurdle for most statisticians is learning to get the data in and out, and selecting subsets of data to work with. The actual statistical commands are usually easy to remember, and to look up if you forget.

The `R` statistical software package

- `R` is a free software implementation of the statistical programming language `Splus` developed and distributed by Bell Labs.
  - You can download it from `http://www.r-project.org/` for Windows and Mac, and some Unix systems. Most Linux and free BSD distributions have prebuilt packages.
  - `R` is not the easiest package to use. `Splus` and `SPSS` are probably much easier with GUIs, while `TSP` and `Shazam` are well-tuned to economics and many business applications. (I use it because it's free software, and offers some extra flexibility I sometimes need. Sorry.)
  - `R` does provide GUI for the Mac (at least Mac OS X 10.5 "Leopard") and Windows; I'll let you know how those work when I've tried them.

A session with the `R` statistical software package

Today we will use `R` to

- Load data from text and `.xls` files
- Print out data sets
- Do some simple regressions and look at the output summaries

Starting R and getting help

- To start R click on the icon, or type R on the command line.
- R help and manuals are all online, distributed with R. Type `help()` for information on the help system, or `help.start()` to bring up a list of resources such as manuals in your web browser (Firefox, Safari, Opera, or IE).
  - Following the trail `An Introduction to R > An Introductory Session > A sample session` and working through the examples is *strongly* recommended.
  - **Note:** R must be running for browser help to work!
- Type `demo()` to get a demonstration of how R works on some more or less real problems.

Loading data into R

- This is something that can be more annoying in R than in more GUI packages like SPSS.
- Start R (see last slide).
- Use `read.table` to read text tables or spreadsheets (including `.xls` and `.csv`).
- For the `.csv` files we use, use the form `db <- read.table("datafile.csv",sep=",",header=TRUE)`.

Hints on `read.table`

- If the form `db <- read.table("datafile.csv",sep=",",header=TRUE)` doesn't work, try reading `help(read.table)`. (Yes, I know it will make your head hurt. Do it anyway, you're a graduate student in training.)
- For some files, `sep` may be a semicolon or tab. Use a text editor (Notepad, Emacs, maybe Word) to look at the file.
- For some files, there may be no variable names, so use `header=FALSE` (or leave out the `header` option).
- The `data` function looks simpler, but that is because it is designed for use with data *pre-packaged for* `R`. This isn't worth the trouble for us.
- If that doesn't help (for most people, it's more pain than it's

worth), ask an expert. *Try classmates first*, that's how they become experts!

An example session: Starting

```
chibi:DataAnalysis steve$ R

R version 2.11.0 (2010-04-22)
Copyright (C) 2010 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> help.start()
starting httpd help server ... done
If the browser launched by '/usr/bin/open' is already running, it is
    *not* restarted, and you must switch to its window.
Otherwise, be patient ...
```

An example session: Load and examine

```
> help(read.table)
> usgdp <- read.table("data/US-GDP-1947.1-2010.1.csv",sep=",",header=TRUE)
> usgdp[0:2]
    Year Quarter
1   1947       1
2   1947       2
3   1947       3
4   1947       4
5   1948       1
    [about 240 lines deleted]
249 2009       1
250 2009       2
251 2009       3
252 2009       4
253 2010       1
```

An example session: Examine parts

```
> usgdp$GDP[1:5]
[1] 237.2 240.4 244.5 254.3 260.3
> usgdp$.Goods[1:5]
[1]  95.6  98.3 100.4 103.5 105.1
> attach(usgdp)
> Year[1:5]
[1] 1947 1947 1947 1947 1948
> GDP[1:5]
[1] 237.2 240.4 244.5 254.3 260.3
```

An example session: A simple regression

```
> result <- lm(Consumption ~ GDP)
> summary(result)

Call:
lm(formula = Consumption ~ GDP)

Residuals:
    Min      1Q  Median      3Q     Max
-181.46  -93.33   24.45   71.70  274.71
```

Regression

- Correlation shows the *statistical strength* of a relationship: how far two variables are from being independent. At a correlation of 1 (or -1), two variables are *perfectly correlated*.
- From a policy standpoint, although correlation between policy and result is necessary (if the result is independent of the policy, there's no point in conducting policy), if the *functional strength* of the relationship is weak, then the policy will be ineffective.
- With imperfect correlation, the relation of changes in one variable to changes in another is uncertain.
- A *regression model* specifies *a combined functional and statistical model, allowing simultaneous estimation of both functional parameters and statistical ones.*

The regression model

- We identify a *dependent (random) variable* Y, and one or more *independent (random) variables* X_1, ..., X_n.
- *Endogenous* is a near synonym for *dependent*. *Explanatory* is a synonym for *independent*, and *exogenous* is a near synonym.
- We assume a *functional relationship* among the variables, $Y = f(X)$, and the *statistical model* that $\epsilon = Y - f(X)$ is a random variable with mean zero ($f(X)$ is an unbiased predictor of Y), and *known* distribution across observations.
- In a data set, this becomes $\epsilon^t = Y^t - f(X^t)$. That is, each observation contains a measurement of Y and of each independent variable X_i. ϵ^t is unobservable, and f is unknown. The problem is to determine f.

The basic linear regression model

- We want to simplify the problem.
- First, we simplify the statistical model by assuming that in the data set, ϵ^1, ..., ϵ^T (T observations on all variables) are *i.i.d.* with mean 0 and variance σ^2.
- Next, we simplify the functional model by assuming that the unknown characteristics are *linear*. That is, the model is that there are coefficients a_T, ..., a_n and $f(X) = \sum_{i=1}^{n} a_i X_i$.
- We can rewrite the model now as

$$Y^t = a_1 X_1^t + \cdots + a_n X_n^t + \epsilon^t, \quad t = 1, \ldots, T.$$

Linear regression

- We often include the *degenerate* or *trivial* random variable $X_1^t \equiv 1$. Then a_1 is the *Y-intercept* of the equation.
  - Statistical packages handle the intercept in different ways. Some require you to specify it explicitly, using a predefined variable (often C or 1). Some provide an option to the regression command to add an intercept term, others provide an option to suppress the intercept term.
- Use of t for "time" is obvious, but it might be that t identifies individuals in a sample, or any other way of collecting observations (*e.g.*, one for each of the prefectures of Japan).

Estimating the parameters

- Our parameters are a_1, ..., a_n, and σ^2.
  - Don't forget σ^2!
  - ϵ is *not* a parameter! It's an unobservable r.v.
- The means of all ϵ^t are *known* to be 0.
- Several strategies for estimation: pick the a_is to
  - Minimize $\sum_{t=1}^{T}(e^t)^2$ where $e^t = Y^t - \sum_{i=1}^{n} a_i X_i^t$ (the *least squares* strategy). This strategy automatically results in $\sum_{t=1}^{T} e_t = 0$.
  - Constrain $\frac{1}{n}\sum_{t=1}^{T} e^t = 0$ and *maximize likelihood* of the configuration of e^ts.

The least-squares formula

- In this model (i.i.d. with symmetric distributions for the ϵ^t), all the plausible strategies lead to the same computation.
- In the *bivariate model with intercept* $Y_t = a + bX_t + \epsilon_t$ (note change of notation! parameters have different letters and the observation index is now a subscript), the formulæ are

$$\begin{aligned}
\hat{b} &= \frac{\sum_{t=1}^{T} x_t y_t}{\sum_{t=1}^{T} (x_t)^2} \\
\hat{a} &= \frac{\sum_{t=1}^{T} Y_t}{T} - \hat{b}\frac{\sum_{t=1}^{T} X_t}{T} \\
\hat{\sigma}^2 &= \frac{\sum_{t=1}^{T} e_t^2}{T-2}
\end{aligned}$$

where $x_t = X_t - \frac{1}{n}\sum_{t=1}^{T} X_t$, $x_t = Y_t - \frac{1}{n}\sum_{t=1}^{T} Y_t$, and $e_t = Y_t - \hat{Y}_t$.

Comments on the formula

- Note the denominator in the formula for $\hat{\sigma}^2$! This is an application of "degrees of freedom." In order to compute e_t, we first must compute $\hat{a}$ and $\hat{b}$, losing 2 degrees of freedom. To get an unbiased estimate of σ^2, we must inflate the sample standard deviation by the factor $\frac{n}{n-2} > 1$.
- The generalization to n variables, with or without intercept, is "a simple matter of linear algebra." We will leave it to the computer.

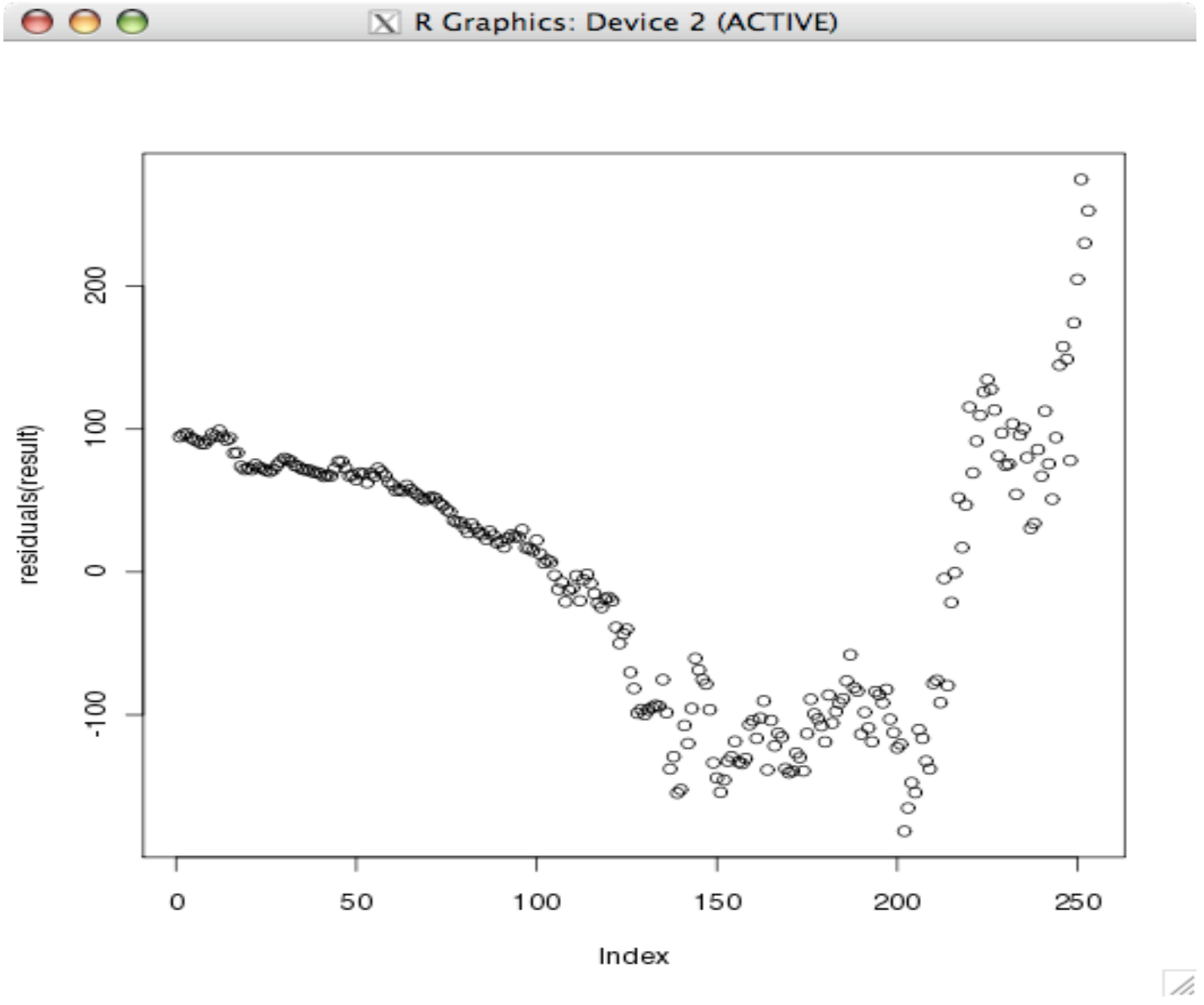
An example session: Regression results

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.042e+02  8.056e+00  -12.93   <2e-16 ***
GDP          6.995e-01  1.321e-03  529.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 91.73 on 251 degrees of freedom
Multiple R-squared: 0.9991,     Adjusted R-squared: 0.9991
F-statistic: 2.802e+05 on 1 and 251 DF,  p-value: < 2.2e-16

> plot(residuals(result))
```

A simple graph



That doesn't look very random!

Factor Analysis

- In regression analysis, we assume we have a good idea explaining the behavior expressed in our data. We represent this explanation as a functional model.
  - Typically, a vector equation $y = f(x)$, *i.e.*,

$$\begin{aligned} y_1 &= f_1(x_1, \ldots, x_k) \\ &\vdots \\ y_n &= f_n(x_1, \ldots, x_k) \end{aligned}$$

  - Sometimes an implicit function: $0 = g(x, y)$.
- In factor analysis, we only have the dependent variables, y, and we want to find a small number of *factors* $x_1, \ldots, x_k$ that explain those variables.

A Simple Example

Consider the following data set, expressed in R:

```
v1 <- c(1,1,1,1,1,1,1,1,1,1,3,3,3,3,3,4,5,6)
v2 <- c(1,2,1,1,1,1,2,1,2,1,3,4,3,3,3,4,6,5)
v3 <- c(3,3,3,3,3,1,1,1,1,1,1,1,1,1,1,5,4,6)
v4 <- c(3,3,4,3,3,1,1,2,1,1,1,1,2,1,1,5,6,4)
v5 <- c(1,1,1,1,1,3,3,3,3,3,1,1,1,1,1,6,4,5)
v6 <- c(1,1,1,2,1,3,3,3,4,3,1,1,1,2,1,6,5,4)
```

- Ignoring the last three elements, `v1`, `v3`, and `v5` are data which are all 1s, except that the 3rd third, the 1st third, and the middle third, resp. are replaced by 3s.
- `v2`, `v4`, and `v6` are `v1`, `v3`, and `v5`, resp., with a little added "noise" (randomness).
- The last three elements ensure nonsingularity.

Correlations for the Simple Example

	v1	v2	v3	v4	v5	v6
v1	1.0000000	0.9393083	0.5128866	0.4320310	0.4664948	0.4086076
v2	0.9393083	1.0000000	0.4124441	0.4084281	0.4363925	0.4326113
v3	0.5128866	0.4124441	1.0000000	0.8770750	0.5128866	0.4320310
v4	0.4320310	0.4084281	0.8770750	1.0000000	0.4320310	0.4323259
v5	0.4664948	0.4363925	0.5128866	0.4320310	1.0000000	0.9473451
v6	0.4086076	0.4326113	0.4320310	0.4323259	0.9473451	1.0000000

- The correlations tell us how closely the variables are related to each other. It should not be surprising that `v1` and `v2` have a very high correlation, and so on.
- Similarly it should be plausible that `v1` and `v3` have a medium correlation.

What Do the Correlations Mean?

- These are artificial data, we know why they are correlated.
- "Eyeballing the numbers," or plotting them on a graph, also makes the relationship clear.
- Sometimes neither is true for "real data."
- We would like an automatic way to "extract" the "causes" of the measured behavior.
- *Factor analysis* of the correlations allows us to do this.

Can We Find Just One "Hidden Cause"?

We ask R to perform a one-factor analysis:

```
factanal(m1, factors = 1)
```

```
Uniquenesses:
   v1    v2    v3    v4    v5    v6
0.773 0.792 0.733 0.795 0.022 0.085


Loadings:
        v1     v2     v3     v4     v5     v6
Factor1 0.476  0.456  0.517  0.453  0.989  0.956

                Factor1
SS loadings       2.800
Proportion Var    0.467
```

```
Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 53.43 on 9 degrees of freedom.
The p-value is 2.43e-08
```

How About Two?

We ask R to perform a two-factor analysis:

```
factanal(m1, factors = 2)
```

```
Uniquenesses:
   v1    v2    v3    v4    v5    v6
0.005 0.114 0.642 0.742 0.005 0.097


Loadings:
        v1     v2     v3     v4     v5     v6
Factor1 0.971  0.917  0.429  0.363  0.254  0.205
Factor2 0.228  0.213  0.418  0.355  0.965  0.928

                Factor1 Factor2
SS loadings       2.206   2.190
Proportion Var    0.368   0.365
```

```
Cumulative Var   0.368   0.733

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 23.14 on 4 degrees of freedom.
The p-value is 0.000119
```

How About Three?

We ask R to perform a three-factor analysis:

```
factanal(m1, factors = 3)
```

```
Uniquenesses:
   v1    v2    v3    v4    v5    v6
0.005 0.101 0.005 0.224 0.084 0.005


Loadings:
        v1    v2    v3    v4    v5    v6
Factor1  0.944    0.905    0.236    0.180    0.242    0.193
Factor2  0.182    0.235    0.210    0.242    0.881    0.959
Factor3  0.267    0.159    0.946    0.828    0.286    0.196


                Factor1 Factor2 Factor3
SS loadings       1.893   1.886   1.797
```

```
Proportion Var   0.316   0.314   0.300
Cumulative Var   0.316   0.630   0.929

The degrees of freedom for the model is 0 and the fit was 0.4755
```

Three with Rotation

We ask R to perform a three-factor analysis:

```
factanal(m1, factors = 3, rotation = "promax")
```

```
Uniquenesses:
   v1    v2    v3    v4    v5    v6
0.005 0.101 0.005 0.224 0.084 0.005


Loadings:

           v1      v2      v3      v4      v5      v6
Factor1                                   0.910   1.033
Factor2   0.985   0.951
Factor3                   1.003   0.867


               Factor1 Factor2 Factor3
```

```
SS loadings      1.903   1.876   1.772
Proportion Var   0.317   0.313   0.295
Cumulative Var   0.317   0.630   0.925

Factor Correlations:
        Factor1 Factor2 Factor3
Factor1   1.000  -0.462   0.460
Factor2  -0.462   1.000  -0.501
Factor3   0.460  -0.501   1.000

The degrees of freedom for the model is 0 and the fit was 0.4755
```

Why no test?

- You may have noticed that there was no report of a hypothesis test for the 3-factor model.
- The reason is that there are no degrees of freedom left (degrees of freedom were zero!)
- Calculating degrees of freedom for the factor analysis is complicated; leave it up to the program.

Is There Really an IQ?

R provides a number of sample datasets and programs, including one on measurements of intellectual ability. But is there a single factor ("IQ") that accounts for all intellectual performance?

```
factanal(factors = 1, covmat = ability.cov)
```

```
Loadings:
        general picture blocks   maze   reading  vocab
Factor1  0.682   0.384   0.502   0.300   0.877   0.849

Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 75.18 on 9 degrees of freedom.
The p-value is 1.46e-12
```

It would appear not!

Multiple Factors in Ability

```
factanal(factors = 2, covmat = ability.cov, rotation = "promax")

Uniquenesses:
general picture  blocks    maze reading   vocab
  0.455   0.589   0.218   0.769   0.052   0.334

Loadings:
        general  picture  blocks    maze    reading   vocab
Factor1   0.364                               1.023    0.811
Factor2   0.470    0.671    0.932    0.508

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 6.11 on 4 degrees of freedom.
The p-value is 0.191
```

In this data set, it seems that there are just two different "kinds" of intelligence, which we could call "geometric" (or "visual") and "verbal". "General intelligence" is related to *both* factors.

Get the data

Due June 20, 11:45.

1. Get the data set `Section1All_csv.csv` from the home page.

   This data set has several sections with different kinds of data. *After reading and thinking about the rest of the problems*, pick one section; using data across sections is a bad idea.

2. Input the data into your statistical package, and print out the data of the section (only!—no fair printing everything and editing the output) you have picked.

   There are two basic ways to accomplish this: create a new data set with exactly the rows and columns you need, or input the whole thing and use the package to pick out "your" variables. Also, many packages prefer that variables be columns and rows be observations, but this sheet has the opposite orientation.

Correlation matrix

3. Generate a correlation matrix for all the variables in your section.

4. Think of some way in which *some* of the variables in your section are related. Refer to scientific theory where possible.

Define and estimate a model

5. Define a regression model for *the variables you picked.*
   (a) Explain why you picked the dependent variable.
   (b) Write down your regression model.
   (c) Estimate the regression model using your statistical package.

Add an unrelated variable

6. Add a random, and therefore unrelated, variable to the model.
   (a) Use Excel or your statistical package to generate a series of random numbers, enough to make a new variable for your data set.
   (b) Add it to the data set, and print out the data set (*i.e.*, your model variables plus the random variable).
   (c) Add the random variable to your model of problem **??** as an explanatory variable, and estimate the new regression model.
   (d) Define and execute a hypothesis test that the new variable is in fact statistically unrelated to the model.

Factor analysis of artificial data

1. Reproduce the factor analysis of six artificial variables done in class using your preferred statistical package.

Factor analysis of real data

2. Using the same data as in the regression problems, conduct a factor analysis on one factor, two factors, *etc.*, until you have "enough" factors.

3. Explain how you know when you have enough. Be quantitative!

Final Examination

- The final examination for this class will be held in **8A108** on Thursday, June 28 from 12:15–15:00.
- Professor Okada will be proctor. No questions will be answered in the exam; if you have a question, he will record it and I will determine later how it impacts the fairness of the exam.
- I plan to include content that was also on the midterm (about 1/3 and no more than 1/2 of the questions), as well as material covered since the midterm (at least 1/2). Conceptual material will be the majority as with the midterm.
- Length will be greater than the midterm, but not 2X as long.

Review Session

- A review session will be scheduled, probably on Friday, June 22, or Monday, June 25, from 5pm-7pm.

  Update: *The review session will meet Friday, June 22, at 8A108, from 16:45 (6th period).*

- Send mail to `data-vote@turnbull.sk.tsukuba.ac.jp` to expression your preference for date. **Update:** *Voting is closed.*

- The mail should have the following content:

  line 1: Your student ID

  line 2: Preferred date/time

  line 3: Two dashes and nothing else: `--`

4 and up: Any other comments about the review session.

- Mail is due by June 14, 11:45 (to allow preparation).

1. Use the spreadsheet `mean-variance.xls` to get three random sequences of numberical scores.
2. Input them into your chosen statistical package, giving them names. Your names should indicate that these series are related to each other.
   - Note that many statistical packages can read Excel spreadsheets. Find out if yours can, and if it does, use that function to create a dataset by copy-paste the random variables into a new spreadsheet.
3. Use your package to print out descriptive statistics for the three variables in that assignment. These statistics should include
   - mean
   - standard deviation

   of each variable, and the
   - correlation matrix

of all variables.

Your package may automatically provide more information as part of a standard "print usual statistics" routine; please include that as well. The correlation matrix sometimes requires a separate command, and sometimes is an option to the general command.