Basic Data Analysis

Stephen Turnbull

Business Administration and Public Policy

Lecture 7: May 23, 2013

Abstract

We continue sampling and estimation of mean and variance. Then we introduce interval estimation and hypothesis testing.

Estimators

- The *process* of (1) computing the mean of the sample and then (2) using it as an estimate of the mean of the population is called an *estimator*.
- An estimator is a process or *algorithm* for making an estimate.
- An *estimator* is a *random variable whose value is used as an estimate of some parameter of interest.*

Sampling

- A *sample* is a set of observations on an "underlying" distribution.
  - The underlying distribution may be an actual population (*e.g.*, our university students).
  - It could be a repeatable random experiment (rolling a die).
  - Or some mixture (typical business problems).
- A *representative sample* is one whose empirical relative frequency distribution is the "same" as the underlying distribution.
  - This must be an approximation, unless we already know the underlying distribution.

Random sampling

- Some samples are inherently based on random events, like rolling a die. There is no physical population to count.
- In the case of a physical population, there are many ways to choose a sample. We can pick the "representative" members.
  - This assumes we know enough to judge which members are representative: but that's what we want to find out!
- If we pick at random, then the population distribution itself determines how likely each member is to be selected for the sample.

Independently, identically distributed

- Usually abbreviated *i.i.d.*

- *Identically distributed* of course just means that we use the same distribution function F for all X_i.

- *Independently distributed* means that for all $i \neq j$, X_i and X_j are independent random variables.

- Recall that X_i and X_j are independent when

$$P[\omega : X_i(\omega) \leq x_i \text{ and } X_j(\omega) \leq x_j] = F(x_i)F(x_j)$$

for all possible values of x_i and x_j (*i.e.*, the values in the support).

Independence and sampling: I

- Consider a jar containing 3 balls, red, white, and blue.
- Suppose we take out a ball, which turns out to be red, and then one which turns out to be blue. What color is the next draw?
- This procedure is called "sampling *without* replacement." The probabilities of the colors *change* with each draw, and therefore the samples are not independent.

Independence and sampling

- Consider our jar containing 3 balls, red, white, and blue.
- Suppose we take out a ball, which turns out to be red, and then *put it back in the jar*. Then take out one which turns out to be blue, and put it back. What can you say about the color of the next draw?
- This procedure is called "sampling *with* replacement." The probabilities of the colors *do not change* with each draw, and therefore the samples are independent.

About "randomness" in sampling

- People often use the word "random" to mean "equal probability," but it doesn't mean that in theory or in practice.
- In theory, as we have seen, even if we don't distinguish among our primitive events $\omega \in \Omega$, we can still assign probabilities to the individual ωs arbitrarily. $P(\omega) = 1/\#\Omega$ is true only if we say it is.
- In practice, there may be latent variables that affect the probabilities of selection. If one of the balls in the jar is coated with a slimy or sticky liquid, it may be less likely to be selected. (If they aren't balls, but rather are cookies, you can probably tell the difference between Oreos and chocolate chip cookies by feel.)
- So even with a "balls in jar" experiment, we need to describe the balls as identical in every way that could affect choice.

When do we use different kinds of sampling?

- With random events, we have no way to control dependence.
- In sampling a univariate variable, we strongly prefer independent observations, and thus for a small population we want random sampling *with* replacement.
- For large populations, random sampling *without* replacement is "close enough" to i.i.d. for our purposes.
  - For observations on people, sampling with replacement is problematic. There's measurement error, so you want to actually ask twice, but then the subject gets annoyed.

Stratified sampling

- For some uses, *stratified sampling* can *improve* representativeness. This relies on *non*-independence!
- Men and women have different distributions of many things. Suppose we have a population which is only 10% female.
- The small number of women in a random sample means statistics for women will be *inaccurate*. Comparisons with men will be *inaccurate*, too.
- The accuracy of the *comparison* can be improved by deliberately constructing a sample with more women than their representation in the population.
  - If the goal of the study is *comparison only*, then having equal numbers of men and women in the sample is best!

Estimating the mean of a distribution, again

- We return to the problem of studying the distribution of heights of the population of students in the university.
- We pick a random sample, which we suppose is therefore representative.
- The mean of the distribution of heights in our sample of students is an estimator for the mean of the heights of in the population.

Random sample and the law of large numbers

- “Random sampling with replacement” guarantees an *identically, independently distributed* sequence of n random variables.
- We use the central limit theorem to determine that the distribution of the mean of the sample (which is a random variable) is a normal distribution, with the same mean as the population, and a variance which is a function of the sample size and the population variance.
- Thus we predict that the mean of the sample will be close to the population mean and that it will not systematically tend to be too large or too small.

The Central Limit Theorem

- The Central Limit Theorem is a very general theorem of probability theory. The version we use is

  > *Let F be a distribution with finite mean μ and finite variance σ^2, and X_i, $i = 1, \ldots, n$ be a sequence of random variables identically and independently distributed according to distribution F. Then $\frac{1}{n}\sum_{i=1}^{n} X_i$ is a random variable whose distribution converges to $N(\mu, \frac{\sigma^2}{n})$ as n becomes large.*

- "Converges" is defined in probability theory; we don't need to know the definition here. *Do* remember that the Central Limit Theorem is an *approximation*.

Estimator bias

- The *bias of $\hat{\mu}$ as an estimator of* μ is defined $\mathcal{E}[\hat{\mu} - \mu]$.
- If an estimator's bias is zero, the estimator is said to be *unbiased*. Otherwise it is *biased*.
- For an unbiased estimator, $\mathcal{E}[\hat{\mu}] = \mu$.
- Although the parameter μ is unknown, we can often still compute bias!

Large sample theory

- Sometimes an estimator $\hat{\mu}$ of μ is biased, but we can show that $\lim_{n\to\infty} P[\omega : \hat{\mu}(\omega) - \mu > \epsilon] = 0$ for any $\epsilon > 0$. Such an estimator is called *consistent*.
- In this case (very) large samples are preferred.
- An *unbiased* estimator is *always* consistent.

Bias of the sample mean

- We are using the *sample mean* $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ as an estimator of the *population mean* μ.

- In a random sample with replacement, each X_i has the same distribution, and therefore the same mean μ, as the population distribution. Thus by linearity

$$\mathcal{E}[\bar{X}] = \mathcal{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n}\mathcal{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$

- In this case, the bias is zero, the sample mean is unbiased:

$$\mathcal{E}[\bar{X} - \mu] = \mathcal{E}[\bar{X}] - \mu = \mu - \mu = 0.$$

Estimating the variance

- The variance (or equivalently, the standard deviation) of the population is obviously an interesting quantity in itself, especially for distributions of known form (such as normal).
- An estimate of variance is essential to estimate the error in other estimates (such as our estimate of the mean).
- It is also essential for *interval estimates* and *hypothesis testing*.

Estimator accuracy

- According to the Central Limit Theorem, $\bar{X}$ has the (approximate) distribution $N(\mu, \frac{\sigma^2}{n})$.
- Let's use the same strategy for estimating σ^2 as we did for μ: take the corresponding variance of the sample.
- This is *non-linear*, so we need to check for bias. Evaluating $\mathcal{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2]$

$$
\begin{aligned}
&= \mathcal{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j)^2] \\
&= \mathcal{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i^2 - \frac{2}{n}X_i\sum_{j=1}^{n}X_j + (\frac{1}{n}\sum_{j=1}^{n}X_j)(\frac{1}{n}\sum_{k=1}^{n}X_k))] \\
&= \mathcal{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i^2 - \frac{2}{n}\sum_{j=1}^{n}X_iX_j + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}X_jX_k)]
\end{aligned}
$$

Evaluating the expectation

- Now we apply linearity and independence for $i \neq j$ and $j \neq k$:

$$
\begin{aligned}
= \quad & \frac{1}{n}\sum_{i=1}^{n}(\mathcal{E}[X_i^2] - \frac{2}{n}(\mathcal{E}[X_i^2] + \sum_{j \neq i}\mathcal{E}[X_i]\mathcal{E}[X_j]) \\
& + \frac{1}{n^2}(\sum_{j=1}^{n}\mathcal{E}[X_j^2] + \sum_{j=1}^{n}\sum_{k \neq j}\mathcal{E}[X_j]\mathcal{E}[X_k]))
\end{aligned}
$$

- We use the property $\mathcal{E}[X_i] = \mu$, and define for convenience $\mu_2 = \mathcal{E}[X_i^2]$ (which makes sense because of identical distributions):

$$
= \quad \frac{1}{n}\sum_{i=1}^{n}(\mu_2 - \frac{2}{n}(\mu_2 + \sum_{j \neq i}\mu^2) + \frac{1}{n^2}(\sum_{j=1}^{n}\mu_2 + \sum_{j=1}^{n}\sum_{k \neq j}\mu^2))
$$

Finishing the evaluation

- Now we collect terms:

$$= \quad \frac{1}{n}\sum_{i=1}^{n}((1-\frac{2}{n}+\sum_{j=1}^{n}\frac{1}{n^2})\mu_2+(\sum_{j=1}^{n}\sum_{k\neq j}\frac{1}{n^2}-\sum_{j\neq i}\frac{2}{n})\mu^2))$$

- Simplify, and restate in expectation and variance terms:

$$\begin{aligned} = \quad & \frac{n-1}{n}(\mu_2-\mu^2)=\frac{n-1}{n}(\mathcal{E}[X_i^2]-(\mathcal{E}[X_i])^2) \\ = \quad & \frac{n-1}{n}\mathcal{V}[X_i]=\frac{n-1}{n}\sigma^2 \end{aligned}$$

- The variance of the sample is a *biased* estimator of the population variance!

Sample variance and standard error

- We define the *sample variance*

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

  which is an unbiased estimator of the population variance, as well as the *sample standard deviation* $s = \sqrt{s^2}$.

- Recall that the variance of the estimator of the mean is $\frac{\sigma^2}{n}$.
  - If we *know* the variance, we use this formula as is, and the *standard error of the estimate* is $\frac{\sigma}{\sqrt{n}}$.
  - If we do not know the variance, but *estimate* it using s^2, then we need to apply the same correction factor as we did to eliminate bias, and the *standard error of the estimate* is $\frac{s}{\sqrt{n-1}}$.

English stinks!

- **Note well:** The *sample variance* ($s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$) is **not** the *variance of the sample* ($\mathcal{E}[(X_i - \mathcal{E}X)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$)!

- The *standard error of the estimate* [of the mean] has two definitions with different formulas, depending on whether we know the true variance, or estimate it with the sample variance.
  - The correction ("multiply by $\frac{n}{n-1}$") is the same in both cases but for somewhat different reasons.

- More technical terms you need to be careful with....

Why the correction factor?

- Recall that when we drew balls from a jar without replacement, the more balls we drew, the better we could predict the next ball. There was less variation, or “freedom,” in the box.
- Similarly, consider this expression from the derivation of the expected value of the variance of the sample:

$$\mathcal{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j)^2] = \mathcal{E}[\frac{1}{n^2}\sum_{i=1}^{n}(nX_i - \sum_{j=1}^{n} X_j)^2].$$

- Note that in the sum over j, there will be an X_i, which cancels one of the n X_is. Thus the estimate actually uses only $n-1$ of the observations, and so is less accurate.

Degrees of freedom

- Since in estimating μ with $\bar{X}$ we use all the data, we say the estimator has n degrees of freedom. When estimating σ^2 with s^s, however, first we must estimate μ with $\bar{X}$, using up one degree of freedom, and leaving only $n-1$ *degrees of freedom* for the estimator for σ^2.

- In general, whether we estimate sequentially (as here) or jointly (as in regression analysis), we count the *degrees of freedom* as $n-(k-1)$ where n is the number of observations, and k is the number of parameters estimated.

How much does the variance vary?

- If you thought to ask "what is the accuracy of the sample variance?", congratulate yourself. You have understood very well!
  - This is the right kind of question.
  - If you are taking statistics (mean, median, or any other), you are doing so to *summarize* varying data; the amount of variation is always important.
- We actually don't normally worry about this, because the sample variance is not easy to interpret, and the variance or standard deviation cannot make more sense than the estimator itself.
- On the other hand, the sample standard deviation is a nonlinear function of the distribution, and calculating its moments is hard.

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 $0.42 \le \alpha \le 0.48$ (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

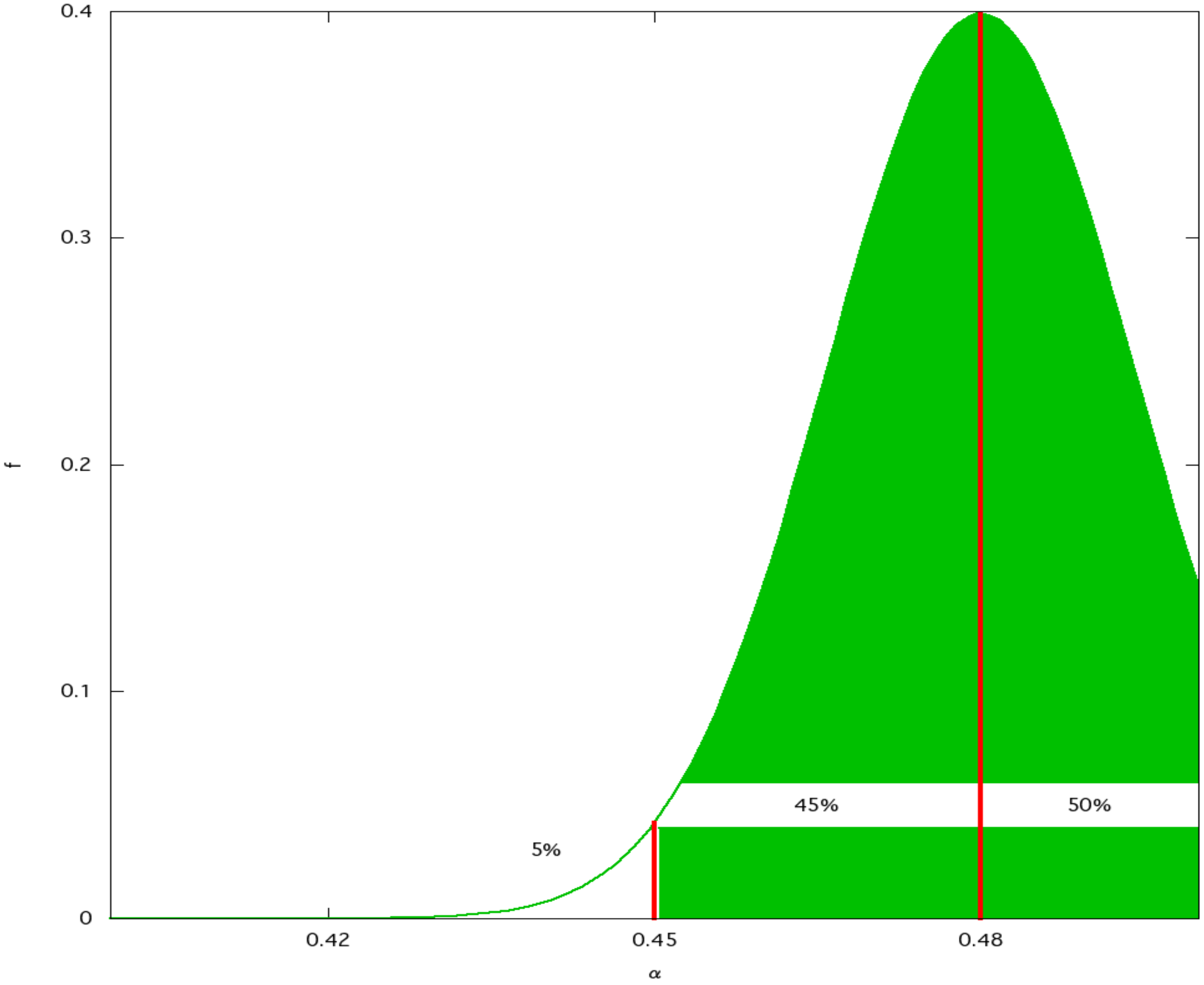
- We quantify "confidence" in probability-*like* terms.
- However, it is *not* a probability. If we estimate the mean by $\bar{X} \pm .03$, the true μ either *is* in the range, or it *is not*. We don't know which is true, but it's *not* random!
- One way to think about it is to try to compute a probability. 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?

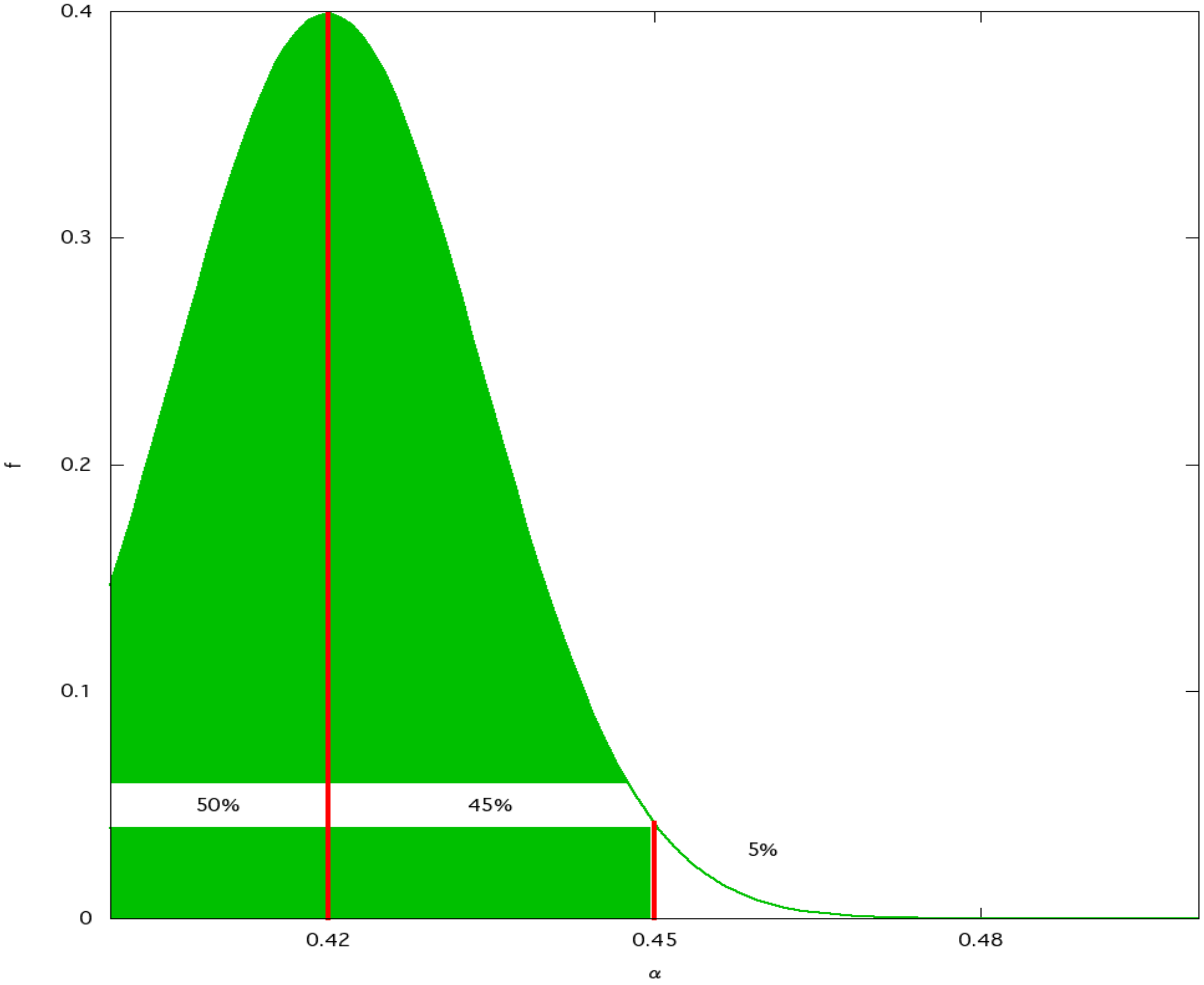
Computing confidence: upper bound

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 $\hat{\alpha}$.



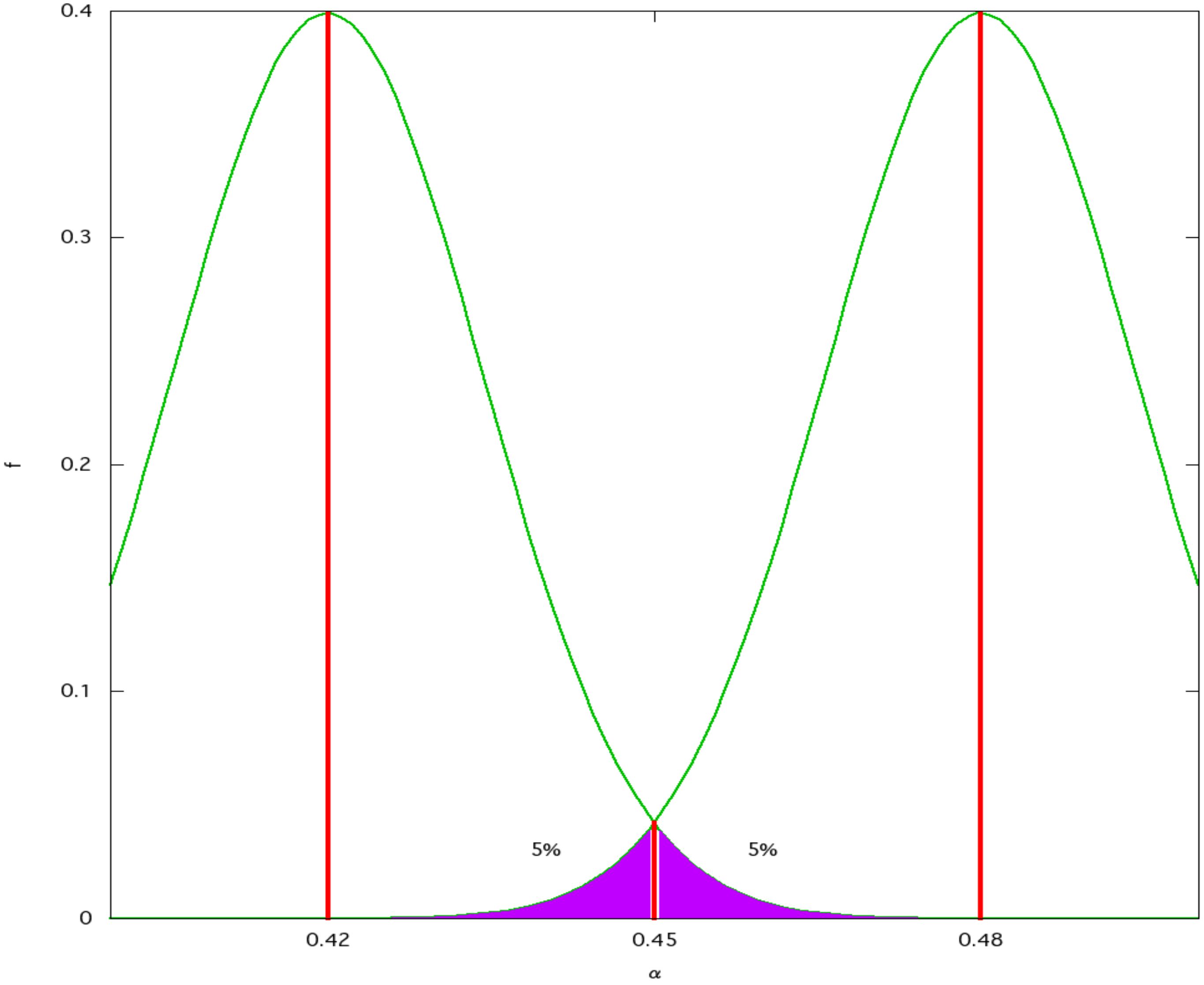
Computing confidence: lower bound

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.



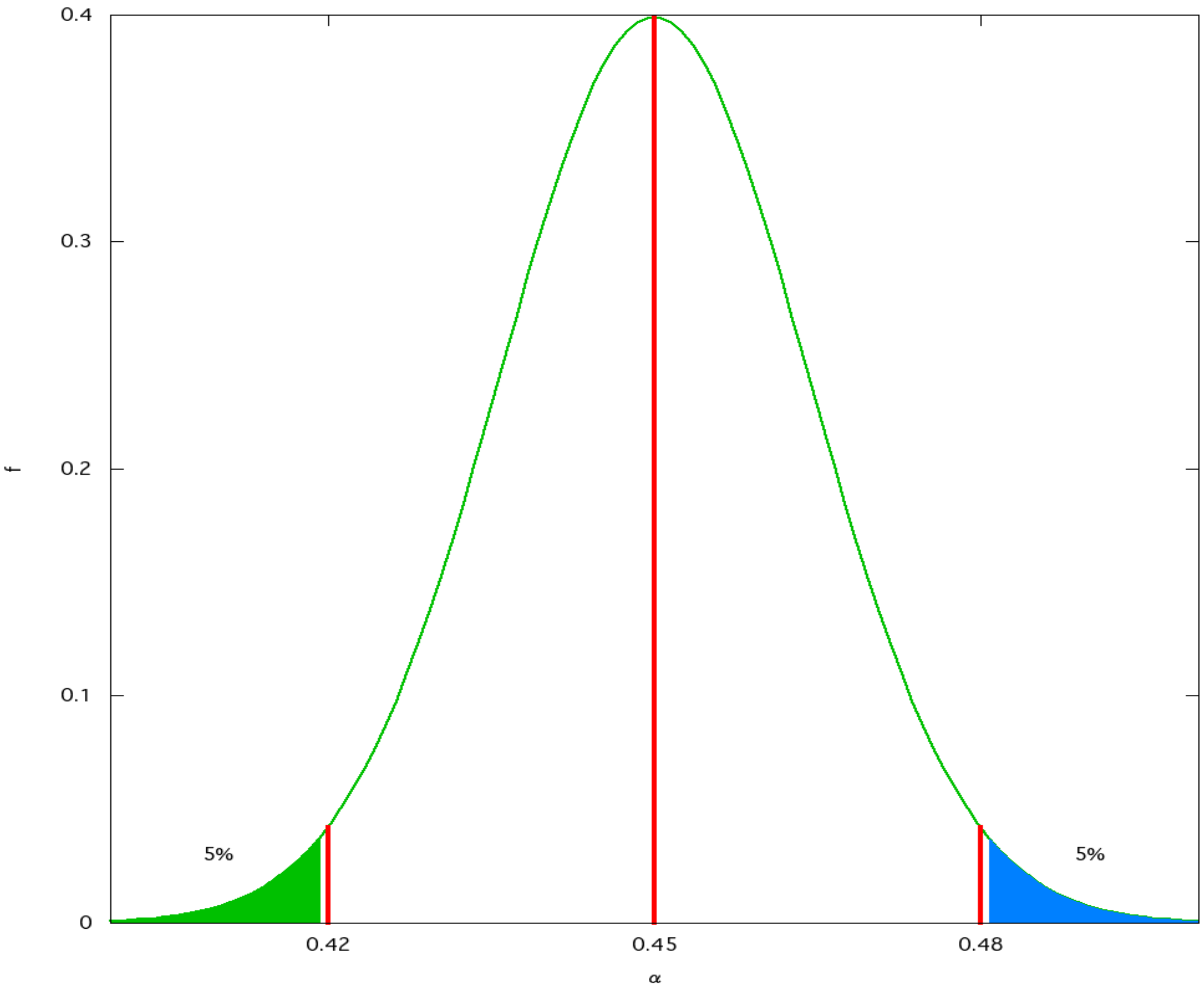
A symmetric interval

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.



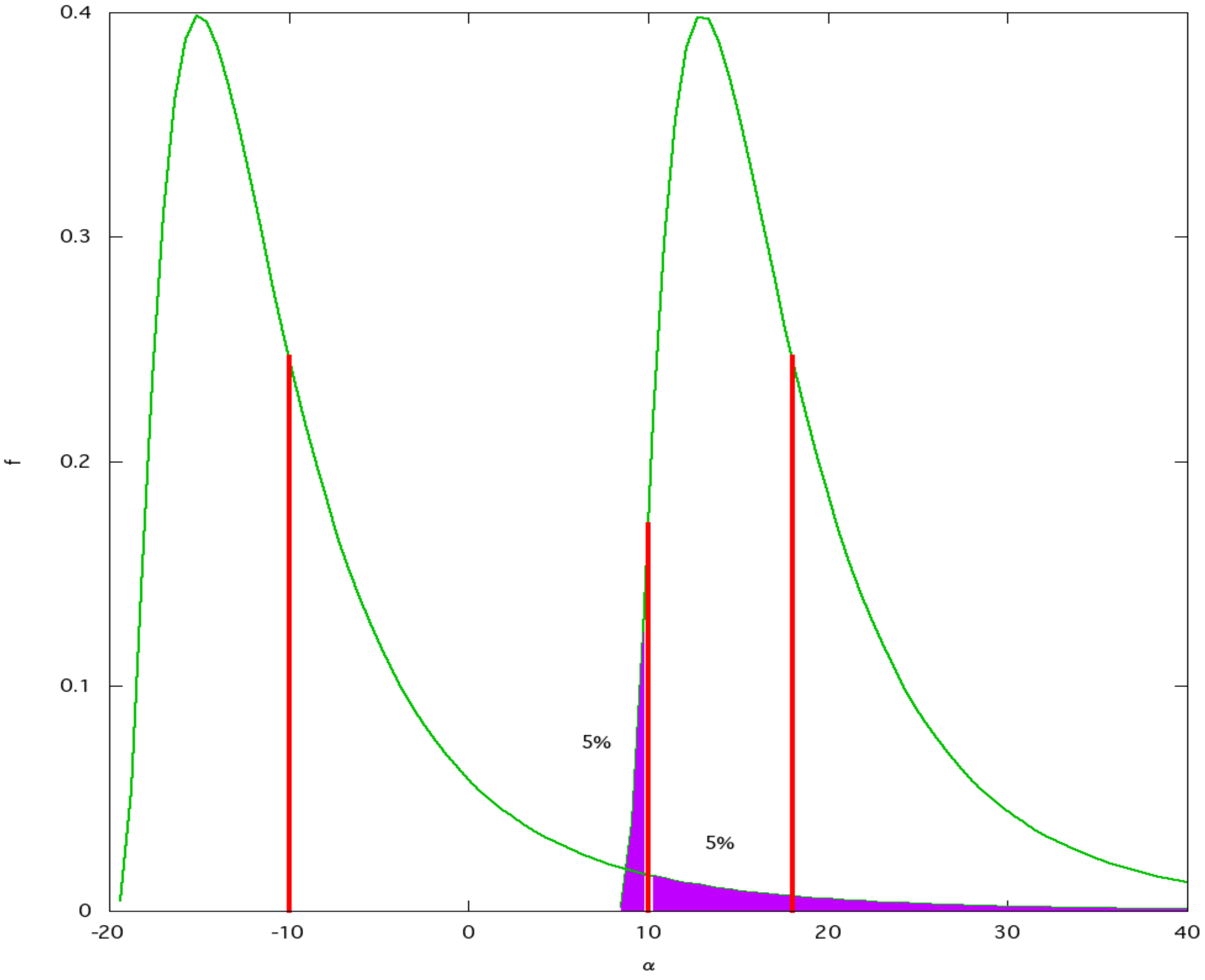
How not to compute a confidence interval

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.

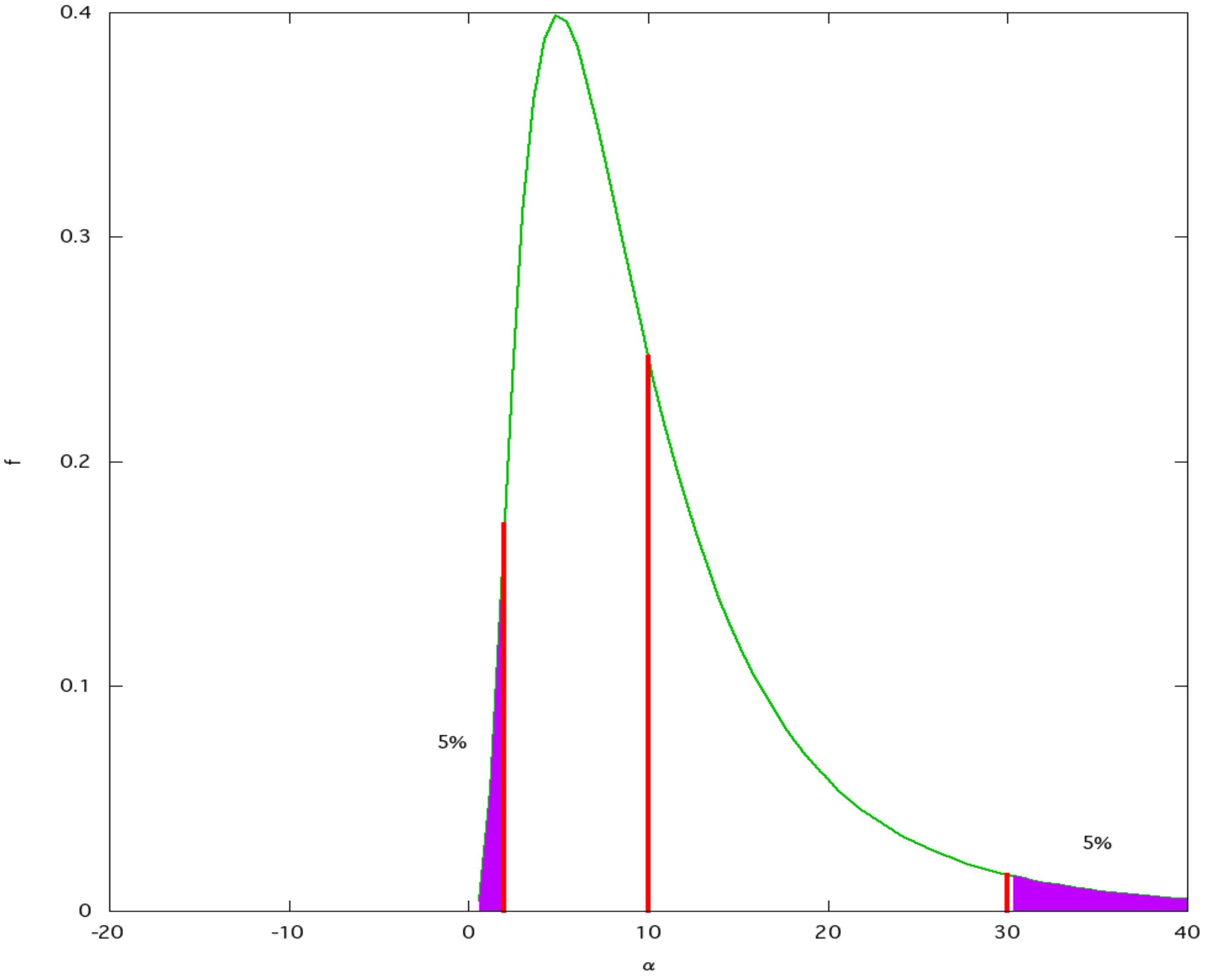


A skewed distribution

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?

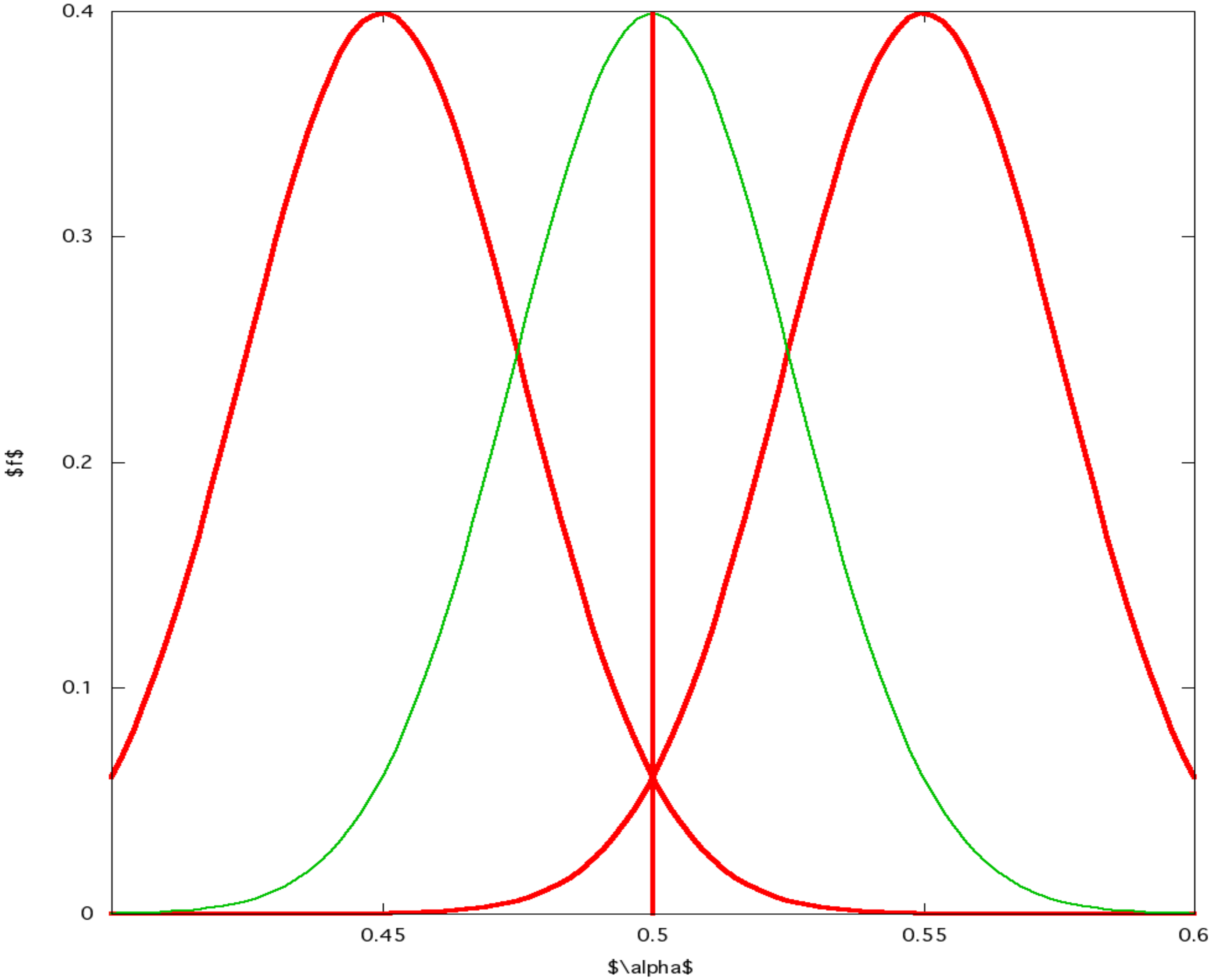
- In Hypothesis 1, “the parties have the same support in the population of voters,” the null hypothesis should be obvious: $H_0 : \alpha = 0.5$.
  - The alternative hypothesis is $H_1 : \alpha \neq 0.5$.
  - Note that H_1 is *almost always* satisfied by the data.
  - But *it cannot be used to compute 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?

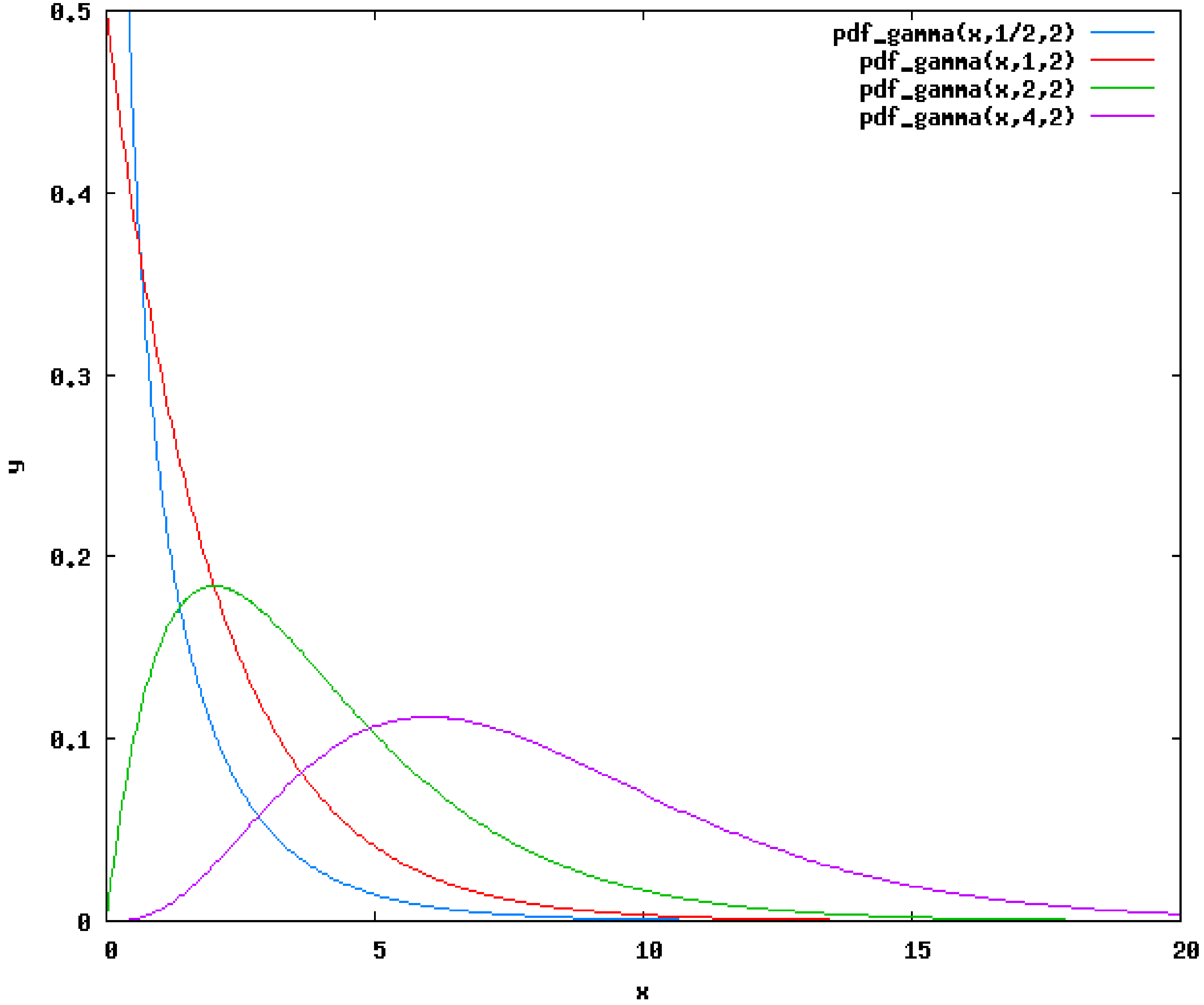
A typical problem of applying statistical theory to real data.

- 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.

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
```

Measurement Project Part II: due May 30, 11:45am

Submit homework to `data-hw@turnbull.sk.tsukuba.ac.jp` by **email**. The due date is May 30, **11:45am**. Submission time is time of receipt by the server.

For this homework, please submit as *plain text* (no wordprocessor or PDF attachments). In other words, "just type" your answer in the email. In the *first line*, include your name, your student ID number, and the words "Measurement 2".

You may write in English or Japanese.

If you wish to ask me questions and get an answer, *write a separate email* to `data-help@turnbull.sk.tsukuba.ac.jp`.

Task

In Part I, you defined an *observation*, that is, a set of variables which will all be measured for each unit of observation. You should have at least five variables in your observation, including one of each of the types (qualitative, ordinal, discrete, and continuous).

1. *Refine* your definition of observation to something where it is practical to get 50 observations. Consider the following:
   - If you propose to survey people (for example, about product satisfaction level), many will refuse to answer.
   - If you propose to survey people at a place of business, you must get the permission of the business's management. **If I receive a complaint about your behavior and you do not have *written* evidence of permission, you will *fail all parts of the measurement assignment.***

- It is probably best to collect data you can easily observe simply by looking and listening.
- If you plan to collect data that comes quickly (*e.g.*, number of cars passing an intersection), you should consider how fast you can record data. This is especially important for cases where you collect multiple variables for a single object.

2. Define *two* relations among your variables you would like to investigate. Each relationship should involve two variables. Describe any cause and effect relationships you believe should hold between them.

You may completely change your plan if you like, but you will need to meet the same conditions as given in the first measurement assignment.

Measurement Project Part III: due June 6, 11:45am

Submit homework to `data-hw@turnbull.sk.tsukuba.ac.jp` by **email**. The due date is June 6, **11:45am**. Submission time is time of receipt by the server.

For this homework, please submit as *plain text* (no wordprocessor or PDF attachments). In other words, “just type” your answer in the email. In the *first line*, include your name, your student ID number, and the words “Measurement 3”.

You may write in English or Japanese.

If you wish to ask me questions and get an answer, *write a separate email* to `data-help@turnbull.sk.tsukuba.ac.jp`.

Task

In Part II, you defined an *observation*, that is, a set of variables which will all be measured for each unit of observation.

1. Collect at least 10 observations to confirm that your data gathering is practical. Having done so, you must be confident that your data is "clean". *Clean* means

   - All observations have all variables recorded.

     In practical statistics, we do frequently encounter missing variables for some observations. However, in this course we do not provide the tools to deal with it.

   - All observations must be accurate.

   Your data should also have a reasonable amount of variation in all variables. If the range of each variable is insufficient, some calculations cannot be done correctly.

2. Pick a statistical package from those available to Keisei students in the lab. It should be capable of doing

   - multiple linear regression analysis
   - factor analysis

   Now

   (a) Enter the numbers from one to ten.

   (b) Give them the variable name "Pikachu."

   (c) Print out the variable Pikachu.

   (d) Compute the mean and standard deviation of Pikachu using your statistical package.

   (e) Save the output to a file named "pikachu.log", and attach the file to your mail for this homework.

   You will be using this package for analysis later.

Measurement Project Part IV: due June 20, 11:45am

Submit homework to `data-hw@turnbull.sk.tsukuba.ac.jp` by **email**. The due date is June 20, **11:45am**. Submission time is time of receipt by the server.

For this homework, please submit as *plain text* (no wordprocessor or PDF attachments). In other words, "just type" your answer in the email. In the *first line*, include your name, your student ID number, and the words "Measurement 4".

You may write in English or Japanese.

If you wish to ask me questions and get an answer, *write a separate email* to `data-help@turnbull.sk.tsukuba.ac.jp`.

Task

In Part III, you did a “pilot study” to show that your project is practical.

1. Complete the study by collecting at least 50 observations.
2. This assignment will be completed later with analytical assignments (statistics for you to compute).