ISTA 116: Statistical Foundations for the Information Age

Sampling Distributions

21 November 2011

Outline

Descriptive vs. Inferential Statistics

- We spent the first several weeks discussing **descriptive statistics**: ways of summarizing and visualizing aspects of data sets.
- This is fine as long as we only care to say something about the specific people, institutions, etc. that we have data about.
- However, the moment we want to *generalize* beyond the data we have in hand to a larger population, we pass to **inferential statistics**.

Probability for Generalization

(Inferential Statistics)

Inferential statistics applies foundational principles from **probability** to the quantities we compute in **descriptive statistics**, to justify generalizations beyond the observed data.

Sample Statistics and Population Parameters

- In many cases, we are interested in some quantity about a **population**, such as a mean, a median, a difference between two group means (or medians), a correlation coefficient, etc.
- These values, applied to the entire population of interest, are called **parameters**.
- However, our data is only a **sample** from a population (ideally a simple random sample); so we don't have access to the true parameter.
- What we do have (in the sample) is a **statistic**.

The Logic of Inferential Statistics

(The Logic of (Classical) Inferential Statistics in a Nutshell)

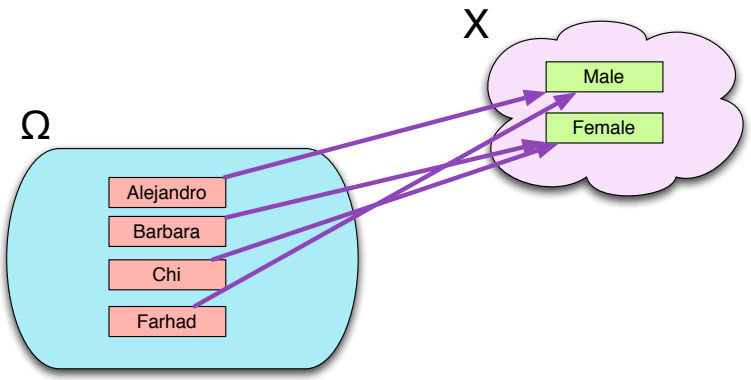
A sample statistic is *evidence* for a hypothesis about a population parameter if the statistic has a value that would be *unlikely* to occur if the hypothesis weren't true.

- To generalize from sample statistics to population parameters, we need to study the probabilistic behavior of sample statistics.

Sampling Values of a Random Variable

- Recall the two defining characteristics of a **simple random sample** from a population:
  1.
  2.
- Each individual in a sample has an associated value for a random variable.

Random Variables



Repeated Sampling

- In a particular sample, we get a particular set of values on the random variable.
- The sample values have a mean, median, variance, etc.
- If we repeated the procedure used to collect the sample, we'd get a new set of values, with a new mean, median, etc.

Sample Statistics are Random!

- Because the sample came from a random process, any statistic has some randomness to it!
- **The sample statistic itself is a random variable** with its own probabilities, mean, variance, etc.
- Q: What's the *sample space* associated with a random statistic?
- A: All possible *samples* having a particular number of observations.

Example: Bernoulli Trials

- We have already seen an example of a sample statistic with a well-defined distribution.
- Suppose I have the hypothesis that smoking rates among UA students have decreased from 30% in 2000. So I randomly sample 10 students for a study, and I classify them into two groups: Smokers and Non-smokers
- What are some things I could say about my sample?

(The Logic of (Classical) Inferential Statistics in a Nutshell)

A sample statistic is *evidence* for a hypothesis about a population parameter if the statistic has a value that would be *unlikely* to occur if the hypothesis weren't true.

Example: Bernoulli Trials

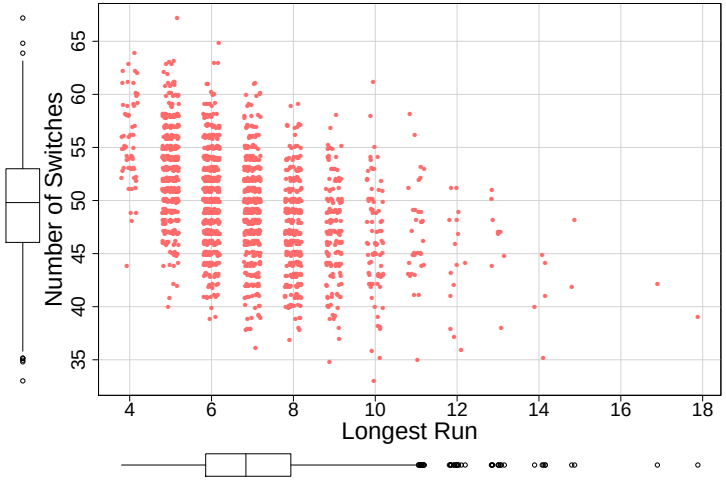
- If my random variable is "number of smokers in the sample", I get one value for each possible random sample of 10 people. What was its distribution in 2000?
- If my hypothesis is that smoking rates have decreased, then I should expect a statistic that has a small probability in this distribution.

Example: Discrete Uniform Trials

- Suppose I roll two dice. Assuming the dice are identical and indepedent, I can think of this as a random sample of two observations from the distribution for one die.
- If I am interested in the sum statistic, what is its sampling distribution for these samples of size two?

An Experiment in Coin-Tossing

Joint Sampling Distribution of Runs and Switches



Sampling Distributions of Sums and Means

- If our sample of size n was produced by simple random sampling from a population, then we have n independent observations of a random variable (or, equivalently, observations from n identical and independent random variables)
- We can take their sum, mean, etc.
- For each set of n people that we sample, we get a different sum, mean, etc.; so the sum is itself a random variable (as is the mean, etc.)
- For the random variables representing the sum and mean in particular, we can say some useful things about their distributions.

Sums of Random Variables

1. If X_1, X_2, $\ldots X_n$ are n random variables, and we define Y to be their sum, then

$$\mu_Y = E(Y) = E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \mu_{X_i}$$

2. If in addition X_1 through X_n are **independent**, then

$$\sigma_Y^2 = Var(Y) = Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} \sigma_{X_i}^2$$

Sums of IID Random Variables

- In the particular case of independent and identically distributed random variables this is even simpler.

(Sums of IID Random Variables)

If $Y = \sum_{i=1}^{n} X_i$ and the X_is are i.i.d.:

1. $\mu_Y = n \times \mu_X$
2. $\sigma_Y^2 = n \times \sigma_X^2$
3. $\sigma_Y = \sqrt{n} \times \sigma_X$

Example: Bernoulli Trials

- Suppose I have the hypothesis that smoking rates among UA students have decreased from 30% in 2000. So I randomly sample 10 students for a study, and I classify them into two groups: Smokers ("1") and Non-smokers ("0")
- The smoking status of each person has a Bernoulli(p) distribution (what's p?). What's its mean? Variance?
- For each sample of 10 people, I can take the sum of these 10 i.i.d. Bernoullis.
- Over all possible samples of 10 people, these sums have a _____ distribution.
- What's its mean, variance and standard deviation?

Example: Discrete Uniform Trials

- Suppose I roll 100 dice. Assuming the dice are identical and indepedent, I can think of this as a random sample of 100 observations from the distribution for one die.
- One die has a mean value of 3.5 and a standard deviation of about 1.71 ($= \sqrt{35/12}$)
- Over all possible rolls of 100 dice, we get a distribution of their sum.
- What's the mean and standard deviation of that distribution?

Means of IID Random Variables

- If we have the sampling distribution of sums, then it's easy to get to the sampling distributions of means: just divide by the sample size.
- What will happen to the mean of a distribution if we divide every value by a constant?
- What will happen to the standard deviation?

Means of IID Random Variables

(Means of IID Random Variables)

If $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n} Y$ (and the X_is are i.i.d.):

1. $\mu_{\bar{X}} = \frac{1}{n} \times n \times \mu_X = \mu_X$
2. $\sigma_Y = \frac{1}{n} \times \sqrt{n} \times \sigma_X = \sigma_X / \sqrt{n}$

- So the mean of the sampling distribution of sample means is the same as the mean of the population distribution, and
- The standard deviation of the sampling distribution of sample means goes down as the sample size increases (but with "diminishing returns", due to the square root)
- Why does this make sense?

Example: Bernoulli Trials

- Suppose I have the hypothesis that smoking rates among UA students have decreased from 30% in 2000. So I randomly sample 10 students for a study, and I classify them into two groups: Smokers ("1") and Non-smokers ("0")
- The smoking status of each person has a Bernoulli(0.3) distribution with mean 0.3 and variance $0.3(1-0.3) = 0.21$.
- For each sample of 10 people, I can take the *mean* of these 10 i.i.d. Bernoullis (in other words, I compute the *proportion* of smokers in the sample).
- What's the mean, variance and standard deviation of the distribution of sample means (i.e., sample proportions)?

Example: Discrete Uniform Trials

- Suppose I roll 100 dice. Assuming the dice are identical and indepedent, I can think of this as a random sample of 100 observations from the distribution for one die.
- One die has a mean value of 3.5 and a standard deviation of about 1.71 ($= \sqrt{35/12}$)
- Over all possible rolls of 100 dice, we get a distribution of their mean.
- What's the mean and standard deviation of that distribution?

Sampling Distributions in General

- In general, it can be quite difficult to find the distribution of a sample statistic exactly, since we have to consider *all possible samples* of a certain size.
- Fortunately, we have two tools for approximation at our disposal:
  1. Simulation
  2. Large sample approximations (e.g., the Central Limit Theorem, which we'll see next week)