Random Sampling

Random samples

A data point is a number (or a color or something). It might be a vector.

A data point is a *realization* of a draw from a *population*. We often assume a hypothetical population which is infinitely countable.

Random samples

Random sample

The vector of random variables $X_1, \ldots, X_n$ is called a *random sample* of size n from the population $f(x)$ if $X_1, \ldots, X_n$ are mutually independent random variables and the marginal pdf of each X_i is the same function $f(x)$.

Alternatively, $X_1, \ldots, X_n$ are called *independent and identically distributed* random variables with pdf $f(x)$. This is commonly abbreviated *iid*.

Joint density

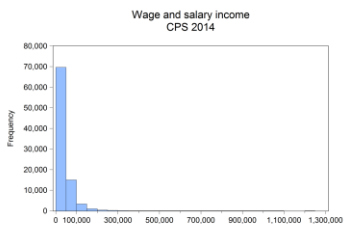
$$f(x_1, \ldots, x_n) = f(x_1) \times \cdots \times f(x_n) = \prod_{i=1}^{n} f(x_i)$$

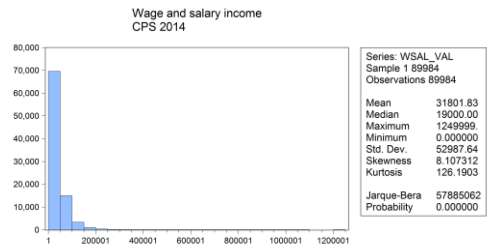
Not everything fits:

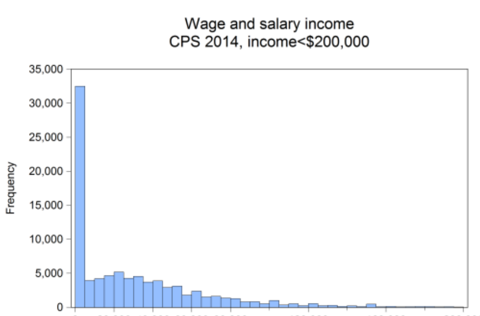
- The draws are not independent.
- The draws are not identical.
- Stratified sample.

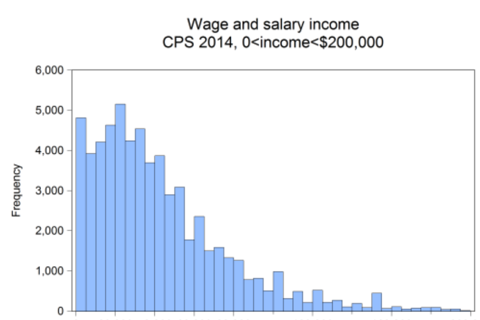
Histogram

- Bar chart where each *bin* gives the count of observations within some range.









Histogram->pdf

Suppose we have B bins with counts c_i and widths w_i. Then for a histogram to represent a pdf we need a normalization constant k to make the cdf end up at 1.

$$1 = \sum_{b=1}^{B} k c_i w_i$$

$$k = \frac{1}{\sum_{b=1}^{B} c_i w_i}$$

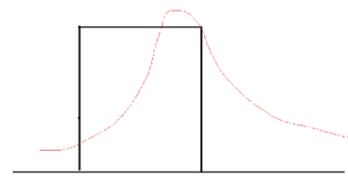
New height is kc_i.

In the case of equal width bins this gives us

$$k = \frac{1}{wn}$$

Kernel smoothing

- Wide bin leads to bias



Narrow bin

A crude estimate of the density function around bin j is

$$\frac{n_j}{nh}$$

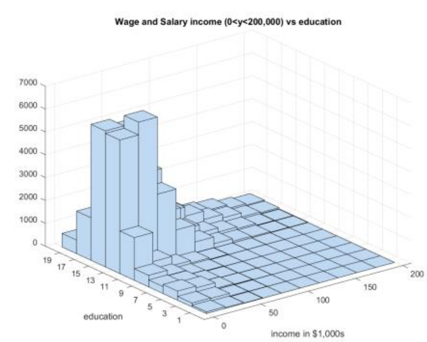
Where h is the width of the bin. n_j is distributed binomial with variance

$$\mathrm{var}(n_j) = np(1-p)$$

Where $p \propto f(\ \ )h$

So the variance of the fraction is

$$\frac{np(1-p)}{(nh)^2} \propto \frac{nf(\ \ )h(1-f(\ \ )h)}{n^2h^2}$$



Finite population

- Two balls in urn: one red, one black

Marginal distribution of a draw is $f(black) = 1/2$, but draws are not independent so two draws is not a "random sample."

Statistic

Let $X_1, .., X_n$ be a random sample of size n from a population and let $T(x_1, ..., x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, .., X_n)$. Then the random variable or random vector $Y = T(X_1, .., X_n)$ is called a *statistic*. The probability distribution of a statistic Y is called the *sampling distribution* of Y.

Examples of statistics

- Height of a bar in a histogram.
- Sample mean
- Sample variance
- Fraction of times a test is rejected in a Monte Carlo experiment.

- Try the following problem. Draw m iid standard normals, $x_i, i = 1, \dots, m$. Compute the statistic

$$T = \frac{1}{m}\sum_{i=1}^{m} I\left(x_i < \Phi^{-1}(.025) \cup x_i > \Phi^{-1}(.975)\right)$$

- Now repeat this experiment n times and show the distribution of T. Do this for $m = 100$ and $n = 1000$. Show what you get empirically as well as what the theoretical answer should be. (Remember that the test statistic is essentially an average of Bernoulli trials.)

Sample mean and variance

The sample mean is

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

And the sample variance is

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

The sample standard deviation is $s = \sqrt{s^2}$.

Theorem 5.2.4: Let $x_1, \ldots, x_n$ be any numbers and $\bar{x} = (x_1 + \cdots + x_n)/n$. Then

a. $\min_a \sum(x_i - a)^2 = \sum(x_i - \bar{x})^2$

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

Proof of part (a). Add and subtract $\bar{x}$ and expand

$$\sum(x_i - a)^2 = \sum\left((x_i - \bar{x}) + (\bar{x} - a)\right)^2$$

$$= \sum(x_i - \bar{x})^2 + 2\sum(x_i - \bar{x})(\bar{x} - a) + \sum(\bar{x} - a)^2$$

The middle term equals zero because $(\bar{x} - a)$ is a constant and $\sum(x_i - \bar{x}) = 0$. The first term is positive. The last term is minimized at zero.

Theorem 5.2.4: Let $x_1, \ldots, x_n$ be any numbers and $\bar{x} = (x_1 + \cdots + x_n)/n$. Then

a. $\min_a \sum(x_i - a)^2 = \sum(x_i - \bar{x})^2$

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

The proof of (b) is

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

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

$$= \sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2$$

Unbiased

$\hat{\theta}$ is an *unbiased estimator* of θ if $\mathrm{E}(\hat{\theta}) = \theta$.

Theorem 5.2.6: Let $x_1, \dots, x_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$.

a. $\mathrm{E}(\bar{x}) = \mu$, $\bar{x}$ is *unbiased*.

b. $\mathrm{var}(\bar{x}) = \frac{\sigma^2}{n}$

c. $\mathrm{E}(s^2) = \sigma^2$, s^2 is *unbiased.*

$\mathrm{E}(\bar{x}) = \mu$, $\bar{x}$ is *unbiased*

$$\mathrm{E}(\bar{x}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mu$$
$$= \mu$$

$\mathrm{var}(\bar{x}) = \frac{\sigma^2}{n}$

$$\mathrm{var}(\bar{x}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\mathrm{var}\left(\sum_{i=1}^{n} x_i\right)$$

$$= \frac{1}{n^2}\left(\sum_{i=1}^{n} \mathrm{var}\, x_i\right) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

$\mathrm{E}(s^2) = \sigma^2$, s^2 is *unbiased.*

$$\mathrm{E}(s^2) = \mathrm{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$
$$= \mathrm{E}\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]\right)$$
$$= \frac{1}{n-1}\left(n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right)$$
$$= \frac{1}{n-1}(n\sigma^2 - \sigma^2) = \sigma^2$$

Biased estimators

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

$$s = (s^2)^{\frac{1}{2}}$$

is a concave function, so by Jensen's inequality

$$\mathrm{E}(s) < [\mathrm{E}(s^2)]^{\frac{1}{2}} = (\sigma^2)^{\frac{1}{2}} = \sigma$$

Example of required sample size calculation

- Without treatment, outcome has mean zero.
- Wish to be able to detect effect of size 2, i.e. reject null of no effect by standard *t*-test.

How large a sample do you need?

Pilot study with no treatment shows outcome $\sim N(0{,}120)$.

Treatment data

$$\hat{\mu} \sim N\left(\mu, \frac{120}{n}\right)$$

Test of null:

$$z = \frac{\hat{\mu}}{\sqrt{120/n}}$$

Reject if $|z| > 1.96$.

Probability of rejection

What is the probability that $|z| > 1.96$ if $\mu = 2$?

$$F_z(-1.96) + \left(1 - F_z(1.96)\right)$$

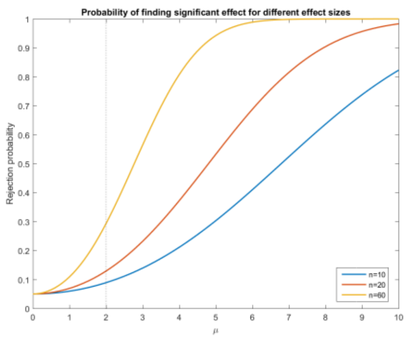
$$z \sim N\left(\frac{\mu}{\sqrt{120/n}}, 1\right)$$

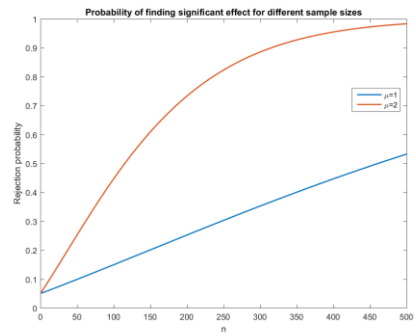
The cdf of z is

$$F_z(\cdot) = \Phi\left(z - \frac{\mu}{\sqrt{120/n}}\right)$$

Probability of rejection is

$$\Phi\left(-1.96 - \frac{\mu}{\sqrt{120/n}}\right) + \left(1 - \Phi\left(1.96 - \frac{\mu}{\sqrt{120/n}}\right)\right)$$





My daughter is going to run a survey in Nigeria. Enumerators are expensive, so each marginal response costs $c = 2500$ naira. A field experiment is planned. A baseline survey was run before any treatment, which indicated that the standard deviation of the variable of interest is $\sigma = 4$ and that the responses are approximately normally distributed. My daughter believes the true effect size is $\mu = 2$. Finding a positive estimated effect size will result in a publication which will have a NPV for her career equal to $50,000. (A reminder that sample means are roughly normally distributed.)

How many surveys should my risk-neutral daughter buy?

