Sampling Distributions

Eli Gurarie

StatR 101 - Lecture 8b
November 15, 2012

November 15, 2012

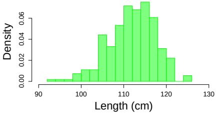
PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON

Review of Definitions

- a **parameter** is a number that describes a "true" distribution
  - μ - mean, σ - standard deviation
  - α, β, γ, etc. in continuous distributions
  - p - probability of a Bernoulli or Binomial
- a **statistic** is *any numerical summary of data*
  - $\overline{X}$ - sample mean
  - s_x - sample standard deviation
  - x_{min}, x_{max}

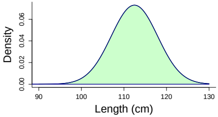
Review of Definitions

Data



Statistics
sample mean: $\overline{x} = 112.5$
sample variance: $s_x^2 = 29.8$
sample s. d.: $s_x = 5.46$

Model



Parameters
theoretical mean: $\mu = 112.5$
theoretical variance: $\sigma^2 = 29.8$
theoretical s.d.: $\sigma = 5.46$

Inference ...

is when you try to say something about a **population** when all you have data on is a **sample**.

- We have a population of interest
- The population has some *true* structure - e.g. a *true* distribution with a *true* parameters (e.g. mean μ, standard deviation σ, p etc.).
- We collect a *random sample* from the population $(X_1, X_2, X_3 ... X_n)$ - because we can't measure the entire population - and calculate a *statistic*,
- We determine how *good* the statistic is at **estimating** the parameter
  - How **accurate** is it? - i.e. how **biased**
  - How **precise** is it? - i.e. how big is the **margin of error** or **confidence interval**.

The most common parameter of interest is ...

$$\mu$$

The best **estimator** of the parameter is ...

$$\overline{X} \approx \mu$$

The big question is ...

How good is $\overline{X}$ at estimating μ?

The reasoning

- The values X_1, X_2, X_3 ... X_n are all *random*, *independent* observations from some distribution $X \sim f(x)$ with some mean value μ. (note that X can be *any* distribution).
- Recall that:
$$\overline{X} = \frac{1}{n}\sum X_i$$
- So: $\overline{X}$ is also a random variable.

Expectation of $\overline{X}$

- Recall the rules of summing and multiplying expectations:

$$\begin{aligned} \mathrm{E}(X+Y) &= \mathrm{E}(X)+\mathrm{E}(Y) \\ \mathrm{E}(aX) &= a\mathrm{E}(X) \\ \mathrm{E}(a+bX) &= a+b\mathrm{E}(X) \end{aligned}$$

- So:

$$\begin{aligned} \mathrm{E}\left(\overline{X}\right) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu = \mu \end{aligned}$$

Expectation of $\overline{X}$

So (using expectation artithmetic):

$$\mathrm{E}\left(\overline{X}\right) = \mu$$

Therefore, we say that: $\overline{X}$ is an **unbiased estimator** of μ, because its *expectation* is exactly the parameter that we are estimating.

Variance of $\overline{X}$

- Here are the complete rules of variance arithmetic:

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(XY) \\
\mathrm{Var}(X-Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(XY) \\
\mathrm{Var}(a+X) &= \mathrm{Var}(X) \\
\mathrm{Var}(bX) &= b^2\mathrm{Var}(X) \\
\mathrm{Var}(aX+bY) &= a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X,Y)
\end{aligned}$$

Brief aside on $\mathrm{Cov}(X,Y)$

The **covariance** of two random variables X and Y is given by:

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathrm{E}\left[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\right] \\
&= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)
\end{aligned}$$

For now, we will not worry about covariances at all, because we will assume that the observations we make are independent.

Rules of expectations and variances (under independence)

- Rules of expectation arithmetic:

$$\begin{aligned} X + Y &= \mathrm{E}(X) + \mathrm{E}(Y) \\ \mathrm{E}(aX) &= a\mathrm{E}(X) \\ \mathrm{E}(a + bX) &= a + b\mathrm{E}(X) \end{aligned}$$

- Rules of variance arithmetic:

$$\begin{aligned} \mathrm{Var}(X + Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) \\ \mathrm{Var}(X - Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) \\ \mathrm{Var}(a + X) &= \mathrm{Var}(X) \\ \mathrm{Var}(bX) &= b^2\mathrm{Var}(X) \\ \mathrm{Var}(aX + bY) &= a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) \end{aligned}$$

Back to the variance of $\overline{X}$

- Recall relevant rules of variance arithmetic:

$$\begin{aligned} \mathrm{Var}(X + Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(XY) \\ \mathrm{Var}(bX) &= b^2\mathrm{Var}(X) \end{aligned}$$

- So:

$$\begin{aligned} \mathrm{Var}\left(\overline{X}\right) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

Expectation and variance of $\overline{X}$

If X is a population with true mean μ and variance σ^2, and X_1, X_2, X_3 ... X_n are observations of that population, and the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ then:

- $\mathrm{E}\left(\overline{X}\right) = \mu$
- $\mathrm{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$

Note that we derived these from the "arithmetic" rules of expectation and variance.

The standard deviation of the sample mean $\sigma_{\overline{X}}$

- The precision of our estimate $\overline{X}$ is determined by the size of our sample:

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$$

- The larger the sample n, the smaller $\sigma_{\overline{X}}$.
- So, because $\mathrm{E}\left(\overline{X}\right) = \mu$ it is an **accurate** (**unbiased**) estimator
- The **precision** of $\mathrm{E}\left(\overline{X}\right)$ depends (as the inverse square root) on n.

Recall the central limit theorem:

Central Limit Theorem (CLT)

If X_1, X_2, X_3 ... X_n are **any, independent, identically distributed (iid)** random variables with mean μ_x and standard deviation σ_x, and

$$Y = \sum_{i=1}^{n} X_i$$

then, as n becomes large

$$Y \sim \mathcal{N}(n\mu_x, n\sigma_x^2)$$

In words: The sum of random variables approximates a normal distribution no matter what the variable is.

A stronger statement about $\overline{X}$

The asymptotic distribution of $\overline{X}$

The distribution of $\overline{X}$ is *approximately normal* with:

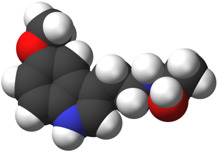
$$\overline{X} \sim \mathcal{N}(\mu, \tfrac{\sigma^2}{n})$$

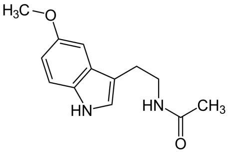
This is a direct consequence of the central limit theorem.

In summary

- Arithmetic rules and central limit theorem let us say anything about the **sampling distribution** of $\overline{X}$.
- So we can solve any probability problem related to $\overline{X}$.

Example with melatonin





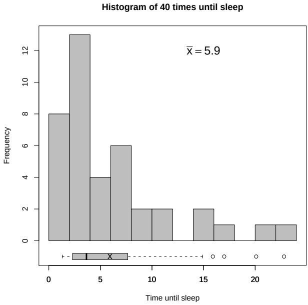
Melatonin

From Wikipedia, the free encyclopedia

Not to be confused with Melanin or Melanotan.

Melatonin /ˌmɛləˈtoʊnɪn/, also known chemically as ***N*-acetyl-5-methoxytryptamine**,[1] is a naturally occurring compound found in animals, plants and microbes.[2][3] In animals, circulating levels of the hormone melatonin vary in a daily cycle, thereby allowing the entrainment of the circadian rhythms of several biological functions.[4]

Example with melatonin



Example with melatonin

- Time to fall asleep for all humans: $\mu = 15$, $\sigma = 10$
- If melatonin has no effect on time to sleep, then

$$X_1, X_2, ..., X_{40} \sim \text{Some Distribution}(\mu = 15, \sigma = 10)$$

- But, by CLT

$$\overline{X} \sim \mathcal{N}\left(\mu = 15, \sigma = \frac{10}{\sqrt{40}}\right)$$
$$\mathcal{N}(\mu = 15, \sigma = 1.58)$$

- So:

$$\begin{aligned} P(\overline{X} \leq 5.9) &= \texttt{pnorm(5.9, mean=15, sd=1.58)} \\ &= \texttt{4.218331e-09} \approx 0 \end{aligned}$$

Another common parameter of interest is ...

$$\sigma^2$$

How good is the **estimator** ...

$$S^2 \approx \sigma^2$$

What about s^2

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

Recall:

$$\sigma^2 = \mathrm{E}\left((X-\mu)^2\right);\ \frac{\sigma^2}{n} = \mathrm{E}\left((\overline{X} - \mathrm{E}\left(\overline{X}\right))^2\right) = \mathrm{E}\left((\overline{X}-\mu)^2\right)$$

$$\begin{aligned}
\mathrm{E}\left(s^2\right) &= \mathrm{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) = \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{E}\left((X_i-\overline{X})^2\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{E}\left((X_i-\mu)^2 - (\mu-\overline{X})^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathrm{E}\left((X_i-\mu)^2\right) - \sum_{i=1}^{n}\mathrm{E}\left((\overline{X}-\mu)^2\right)\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\sigma^2 - \sum_{i=1}^{n}\frac{1}{n}\sigma^2\right) \\
&= \frac{1}{n-1}(n\sigma^2 - \sigma^2) = \frac{n-1}{n-1}\sigma^2 = \sigma^2
\end{aligned}$$

What about s^2

So:

$$\mathrm{E}\left(s^2\right) = \sigma^2$$

This means: s^2 is an **unbiased estimator** of σ^2.

The $n-1$ in the denominator in $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$ is called: **the degrees of freedom**.

(more about this tricky concept later).

Take home message on Sampling Distributions

- Certain *sample statistics* estimate *population parameters*.
  - for example: $\overline{X}$ estimates μ.
- Those statistics are random variables, because *every time you sample you get a different outcome!*
- Probability theory tells us the distribution of these random variables.
  - for example: $\overline{X} \sim \mathcal{N}\left(\mu_x, \frac{\sigma^2}{n}\right)$
- This allows us to infer something sophisticated about the population!

Inference example: Shaq

Do we *know* that Shaq's probability of getting a free throw is 37.4%?

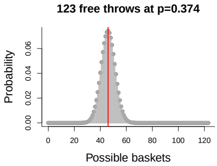
No! We estimate it based on him making 46 shots out of 123 tries.

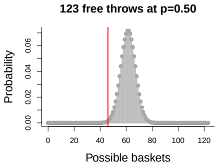
$$\begin{aligned} \widehat{p} &= 46/123 = 0.374 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i \end{aligned}$$

But is it possible that he gets 46/123 shots with a true $p = 0.5$ (for example)?

Possible probability mass functions of 123 baskets

Is it possible that he gets 46/123 shots with a true $p = 0.5$?





Yes! (Almost) anything is *possible*.

Question

- How good is $\widehat{p} = \overline{X}$ as an **estimator** of the true p?

A more "formal" statement

- What is the **sampling distribution** of $\widehat{p} = \overline{X}$

Expectation of $\widehat{p}$

Assumption: $X \sim$ Bernoulli(p)and X are i.i.d. (independent and identically distributed).

$$
\begin{aligned}
\mathrm{E}\left(\widehat{p}\right) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} p \\
&= \frac{1}{n} np = p
\end{aligned}
$$

Variance of $\widehat{p}$

Assumption: $X \sim$ Bernoulli(p)and X are i.i.d. (independent and identically distributed).

$$
\begin{aligned}
\mathrm{Var}\left(\widehat{p}\right) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{n}{n^2}\mathrm{Var}\left(X_i\right) \\
&= \frac{p(1-p)}{n}
\end{aligned}
$$

Let's consider some examples:

- Shaq shoots two free throws, and makes 1 of 2.

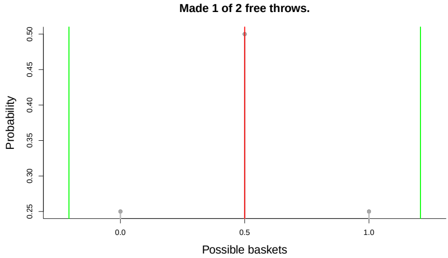
$$\begin{aligned} \widehat{p} &= (0+1)/2 \\ \text{Var}\,(\widehat{p}) &= (1/2)(1/2)(1/2) = 0.125 \\ \text{s.d.}(\widehat{p}) &= \sqrt{0.125} = 0.3535 \end{aligned}$$

So the estimate of $\widehat{p} = 0.5 \pm 0.707$ is not very precise.

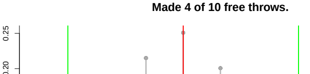
Important note

- People often report estimates as $\widehat{\mu} \pm 2\widehat{\sigma}$
- The range: $(\widehat{\mu} - 1.96\,\widehat{\sigma}, \widehat{\mu} + 1.96\,\widehat{\sigma})$ is referred to as: the **95% Confidence Interval**

Let's consider some examples:

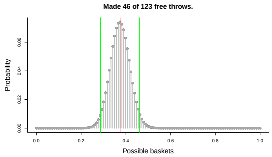


$\widehat{p} = 0.5 \pm 0.707$, C.I. $=$ (-0.21,



Inference question

Given that Shaq shot 46/123 free throws, could he have been a 50% free throw shooter?



- Parameter estimation: $\widehat{p} = 0.374 \pm 0.087$, C.I. $= (0.29, 0.46)$.
- 50% is outside of the **Confidence Interval** of our estimate ... so very, very unlikely ($< 5\%$ chance).