DS-GA 1002 Lecture notes 4 October 2, 2015

Statistics

1 Estimation of population parameters

The goal of statistics is to extract information from data and evaluate the uncertainty of this information quantitatively. In this section we consider the problem of estimating a **population parameter**, which is just a fancy way of calling a fixed quantity of interest such as the average height in New York, the voting intentions of the people in the UK in the Scottish referendum, the variance of tree height in Yosemite National Park, etc. We will consider estimates obtained by sampling a subset of individuals from the population and study how to quantify their accuracy probabilistically.

1.1 Random sampling

Samples from a subset of individuals belonging to a larger group can be interpreted as a sequence of random variables. We often make the assumption that these random variables are **independent** and **identically distributed** (iid).

Definition 1.1 (Sequence of iid random variables)**.** *A sequence of random variables* $X_1, X_2, \ldots$ *belonging to the same probability space is independent, identically distributed if any subset of* n *random variables of the sequence is mutually independent and every* X_i *has the same marginal distribution.*

Note that we are *not* suggesting that the quantity of interest is random; it is the measurement process that is random. In the following example we show how an iid sequence of data is obtained when we sample from a population, even though the actual value for each individual is deterministic.

Example 1.2 (Sampling from a large population)**.** Assume that there are n people in a population and we are interested in measuring their height. To simplify matters, let us assume that there are m possible heights $\{h_1, h_2, \ldots, h_m\}$ and let us denote by n_j the number of people whose height is equal to h_j, $1 \leq j \leq m$. We gather the data by sampling the n individuals uniformly *with replacement*. Note that this means that a person can be chosen several times. The pmf of each sample is

$$p_{X_i}(h_j) = P\left(\text{Chosen person has height } h_j\right) = \frac{n_j}{n}. \tag{1}$$

Every sample has the same pmf, so the data are identically distributed. By assumption, each sample is independent of the others. In fact, this is why we sample with replacement; otherwise the pmf would depend on previous samples. The sequence of data is iid.

At first glance it might seem strange to sample at random from a population. One would think that there should be more efficient ways of obtaining a representative subset. However, as we will see below, sampling randomly allows us to establish precise quantitative guarantees on the accuracy of the information that we extract from the samples. This is not the case if we sample deterministically, as we cannot be sure that we are not systematically ignoring certain subsets of the population.

1.2 The sample mean and variance

A statistical **estimator** is a function of the available data that approximates a parameter of interest. To fix ideas, assume that we want to estimate the mean of a certain quantity from some samples taken from a large population, as in the setting described by Example 1.2. We assume that we have an iid sequence $X_1, X_2, \ldots$. Our aim is to estimate the *true* mean $\mu = \mathrm{E}(X_i)$, which is the same for all X_i because the sequence is identically distributed. A reasonable estimator is the **sample mean**.

Definition 1.3 (Sample mean). *Let $X_1, X_2, \ldots$ be a sequence of random variables belonging to the same probability space. The sample mean is defined as*

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{2}$$

Since we are assuming that an estimator is a function of random samples, we cannot evaluate its performance deterministically. Instead, we can consider its **mean squared error**, which is the expected square difference between the estimator and the parameter of interest.

Definition 1.4. *The mean squared error between an estimator A and the corresponding population parameter θ is*

$$MSE(A) := \mathrm{E}\left((A - \theta)^2\right). \tag{3}$$

The mean squared error can be decomposed into a **bias** term and a **variance term**. The bias term is the difference between the parameter of interest and the expected value of the estimator. The variance term corresponds to the variation of the estimator around its expected value.

Lemma 1.5 (Bias-variance tradeoff)**.** *For any estimator A of a population parameter μ,*

$$MSE(A) = \underbrace{\mathrm{E}\left((A - \mathrm{E}(A))^2\right)}_{\text{variance}} + \underbrace{\mathrm{E}\left((\mathrm{E}(A) - \theta)^2\right)}_{\text{bias}}. \tag{4}$$

Proof. The lemma is a direct consequence of linearity of expectation. □

An estimator is said to be **unbiased** if its bias is zero, i.e. its mean is exactly equal to the parameter of interest. The sample mean is an unbiased estimator as long as all the random variables in the sequence have the same mean.

Lemma 1.6. *If* $\mathrm{E}(X_i) = \mu$ *for all the random variables in a sequence then*

$$\mathrm{E}\left(\bar{X}_n\right) = \mu. \tag{5}$$

Proof. The lemma follows immediately from linearity of expectation. □

The estimator is equal to the parameter of interest on average, which is reassuring. A corollary to the following theorem, a generalization of Corollary 1.19 from Lecture Notes 3, allows to do bound its variance, and equivalently its MSE by Lemma 1.5. The theorem is proved in Section A of the appendix.

Theorem 1.7. *If* $X_1, \ldots, X_n$ *are uncorrelated random variables belonging to the same probability space, then*

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i). \tag{6}$$

Corollary 1.8 (Variance of sample mean)**.** *If* $X_1, \ldots, X_n$ *is a sequence of uncorrelated random variables with the same mean* μ *and variance* σ^2*, then*

$$\mathrm{Var}\left(\bar{X}_n\right) = \frac{\sigma^2}{n}. \tag{7}$$

Proof.

$$\mathrm{Var}\left(\bar{X}_n\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \tag{8}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) \tag{9}$$

$$= \frac{\sigma^2}{n}. \tag{10}$$

□

Another example of an estimator obtained by averaging over a subset of samples is the **sample variance**, which approximates the variance of a sequence of variables.

Definition 1.9 (Sample variance)**.** *Let $X_1, X_2, \ldots$ be a sequence of random variables belonging to the same probability space. The sample variance is defined as*

$$S^2 := \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2 \tag{11}$$

The careful reader might be wondering why the term in the denominator of (11) is equal to $n-1$ and not n. Intuitively, the reason is that we are using the sample mean instead of the mean. Since the sample mean is computed using the same samples that we are using to compute the sample variance, it will tend to be slightly closer to those samples than the true mean. In fact, the denominator must be equal to $n-1$ for the estimator to be unbiased.

Lemma 1.10. *If $X_1, X_2, \ldots$ are uncorrelated and share the same mean μ and variance σ^2 then*

$$\mathrm{E}\left(S^2 \right) = \sigma^2. \tag{12}$$

The lemma is proved in Section B of the appendix.

An important question is whether an estimator becomes more accurate as it incorporates more data. If an estimator becomes arbitrarily accurate with high probability as the number of data increases, then we say that it **converges** to the quantity of interest. We will make this statement precise in the following sections, which introduce the Law of Large Numbers and the Central Limit Theorem.

1.3 The Law of Large Numbers

Convergence for a deterministic sequence is simple to define: $x_n \to x$ as $n \to \infty$ if x_n stays arbitrarily close to x for large enough n.[1] In this sense, it is easy to define convergence for a *realization* of the random sequence X_n, but what does it mean for the random sequence itself to converge to a random variable X? There are several possible answers to this question.

As a first option, we consider the probability of the event $\{X_n \text{ converges to } X\}$. If this probability equals one then we say that X_n converges to X **with probability 1**.

Definition 1.11 (Convergence with probability 1)**.** *Let $X_1, X_2, \ldots$ be a sequence of random variables. The sequence converges with probability 1 to another random variable X belonging to the same probability space $(\Omega, \mathcal{F}, \mathrm{P})$ if*

$$\mathrm{P}\left(\left\{ \omega \mid \omega \in \Omega, \quad \lim_{n \to \infty} X_n(\omega) = X(\omega) \right\} \right) = 1. \tag{13}$$

[1] More formally, for any $\epsilon > 0$ there is an n_0 such that for all $n > n_0$ $|x_n - x| < \epsilon$.

Since X_n and X belong to the same probability space, each ω in the sample space Ω fixes a realization $X_n(\omega)$ and $X(\omega)$ for which we can evaluate convergence in a deterministic sense. Recall that the sample space Ω may be very complicated and we don't usually manipulate it explicitly.

In a nutshell, convergence with probability one means that almost all sequences converge deterministically as $n \to \infty$ to the corresponding realization of the limit random variable X. An alternative viewpoint is to fix n and consider how close X_n is to X at that particular n *probabilistically.* A measure of the distance between two random variables is the mean square of their difference (recall that if $\mathrm{E}\left((X-Y)^2 = 0\right)$ then $X = Y$ with probability one by Chebyshev's inequality). The mean square deviation between X_n and X is a deterministic quantity (a number), so we can evaluate its convergence as $n \to 0$. If it converges to zero then we say that the random sequence **converges in mean square**.

Definition 1.12 (Convergence in mean square)**.** *Let $X_1, X_2, \ldots$ be a sequence of random variables. The sequence converges in mean square to another random variable X belonging to the same probability space if*

$$\lim_{n\to\infty} \mathrm{E}\left((X - X_n)^2\right) = 0. \tag{14}$$

Alternatively, we can consider the probability that X_n is separated from X by a certain fixed $\epsilon > 0$. If for any ϵ, no matter how small, this probability converges to zero as $n \to \infty$ then we say that the random sequence **converges in probability**.

Definition 1.13 (Convergence in probability)**.** *Let $X_1, X_2, \ldots$ be a sequence of random variables. The sequence converges in probability to another random variable X belonging to the same probability space if for any $\epsilon > 0$*

$$\lim_{n\to\infty} \mathrm{P}\left(|X - X_n| > \epsilon\right) = 0. \tag{15}$$

Note that as in the case of convergence in mean square, the limit in this definition is deterministic, as it is a limit of probabilities, which are just real numbers.

As a direct consequence of Markov's inequality, convergence in mean square implies convergence in probability.

Theorem 1.14. *Convergence in mean square implies convergence in probability.*

Proof. We have

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{P}\left(|X - X_n| > \epsilon\right) &= \lim_{n\to\infty} \mathrm{P}\left((X - X_n)^2 > \epsilon^2\right) \qquad (16)\\
&\leq \lim_{n\to\infty} \frac{\mathrm{E}\left((X - X_n)^2\right)}{\epsilon^2} \quad \text{by Markov's inequality} \qquad (17)\\
&= 0, \qquad (18)
\end{aligned}$$

if the sequence converges in mean square. □

When an estimator converges to the true value of the quantity of interest as the number of data grows asymptotically, then the estimator is said to be **consistent**. Let us now return to the problem of estimating the true mean of a sequence of iid random variables: the sample mean is a consistent estimator as long as the random variables in the sequence have finite variance.

Theorem 1.15 (The sample mean is consistent)**.** *If $X_1, X_2, \ldots$ is an iid sequence with bounded variance, such that $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, the sample mean $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ converges in mean square to μ.*

Proof.

$$\lim_{n\to\infty} \mathrm{E}\left(\left(\bar{X}_n - \mu\right)^2\right) = \lim_{n\to\infty} \mathrm{Var}\left(\bar{X}_n\right) \quad \text{by Lemma 1.6} \tag{19}$$

$$= \lim_{n\to\infty} \frac{\sigma^2}{n} \quad \text{by Corollary 1.8} \tag{20}$$

$$= 0. \tag{21}$$

□

Corollary 1.16 (Weak Law of Large Numbers)**.** *Under the assumptions of Theorem 1.15 the sample mean converges in probability to the mean of the distribution μ.*

Proof. The result follows from Theorem 1.14. □

Remark 1.17. *Under the assumptions of Theorem 1.15 the sample mean also converges to the true mean with probability one, a result called the Strong Law of Large Numbers. Convergence with probability one is a stronger statement than convergence in probability.*[2] *A proof of the Strong Law of Large Numbers is beyond the scope of these notes. We refer the interested reader to more advanced texts in probability theory.*

Figure 1 shows a histogram of the heights of a group of 25 000 people[3], which we can use to represent a large population. In Figure 2 we plot three different realizations of the sample mean. As predicted by the Law of Large Numbers, the sample mean quickly converges to the true mean.

[2] This is not the case for convergence in mean square; convergence with probability one does not imply convergence in mean square and convergence in mean square does not imply convergence with probability one.

[3] The data set can be found here: `wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights`.

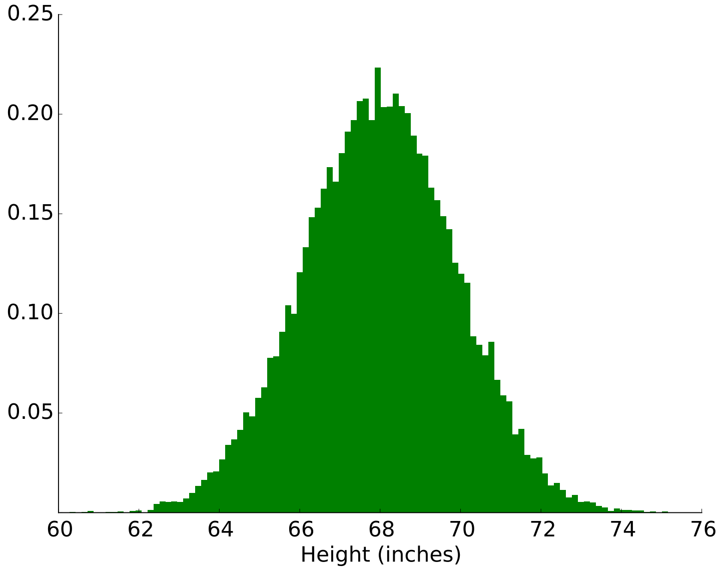


Figure 1: Histogram of the heights of a group of 25 000 people.

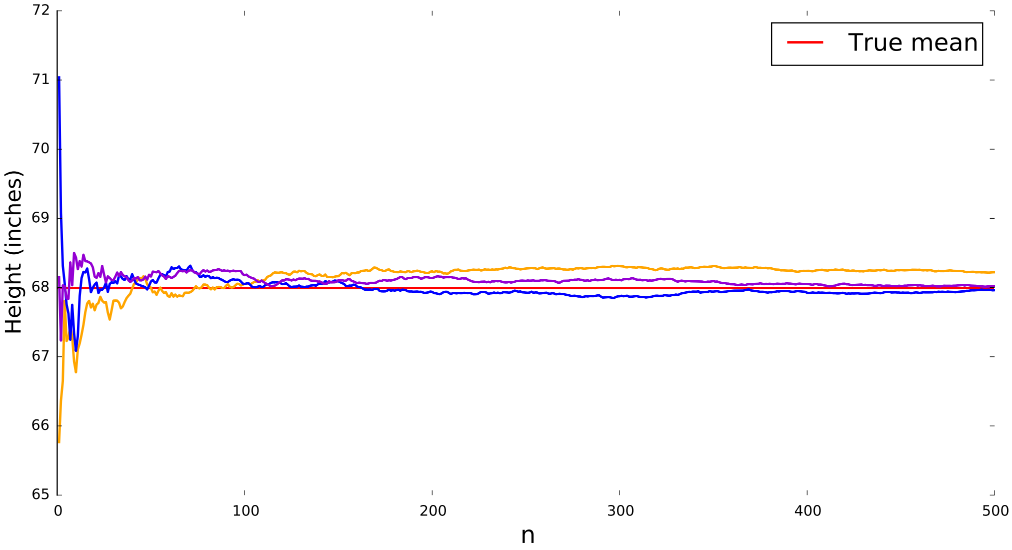


Figure 2: Different realizations of the sample mean for iid samples from the distribution in Figure 1.

1.4 Central Limit Theorem

In the previous section we established that the sample mean is a consistent estimator of the true mean of a sequence of iid random variables. In this section, we will characterize the distribution of the estimator as the sample size n increases. For this purpose, we introduce the notion of convergence in distribution.

Definition 1.18 (Convergence in distribution)**.** *Let $X_1, X_2, \ldots$ be a sequence of random variables. The sequence converges in distribution to another random variable X with cdf F_X belonging to the same probability space if*

$$\lim_{n \to \infty} p_{X_n}(x) = p_X(x) \quad \textit{for all } x \in R_X \tag{22}$$

for discrete X, where R_X is the range of X, or if

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x) \quad \textit{for all } x \in \mathbb{R} \tag{23}$$

for continuous X (note that the X_n are not necessarily continuous!).

Note that convergence in distribution is a much weaker notion than convergence in mean square or in probability. If a sequence $X_1, X_2, \ldots$ converges to a random variable X in distribution, this only means that as n grows the cdf (and consequently the pmf or the pdf, if it exists) of X_n is close to the cdf of X, not that the values of the two random variables are close.

The Central Limit Theorem states that if an iid sequence has the same mean and variance, and the variance is bounded, then the distribution of the sample mean becomes Gaussian in the limit. This holds for any sequence, no matter what its marginal distribution looks like.

Theorem 1.19 (Central Limit Theorem)**.** *If $X_1, X_2, \ldots$ is an iid sequence with bounded variance, such that $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, the sample mean $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ converges in distribution to a Gaussian random variable with mean μ and variance σ^2/n.*

Proof. The proof of this remarkable result is beyond the scope of these notes, but can be found in any advanced text on probability theory. However, we would still like to provide some intuition as to why the theorem holds. The following lemma, proved in Section C of the appendix, characterizes the joint distribution of sums of independent random variables.

Lemma 1.20 (Sum of random variables)**.** *Let X and Y be two independent random variables and let $Z := X + Y$. If X and Y are discrete*

$$p_Z(z) = \sum_{u=-\infty}^{\infty} p_X(u)\, p_Y(z-u) \tag{24}$$

$$= (p_X * p_Y)(z). \tag{25}$$

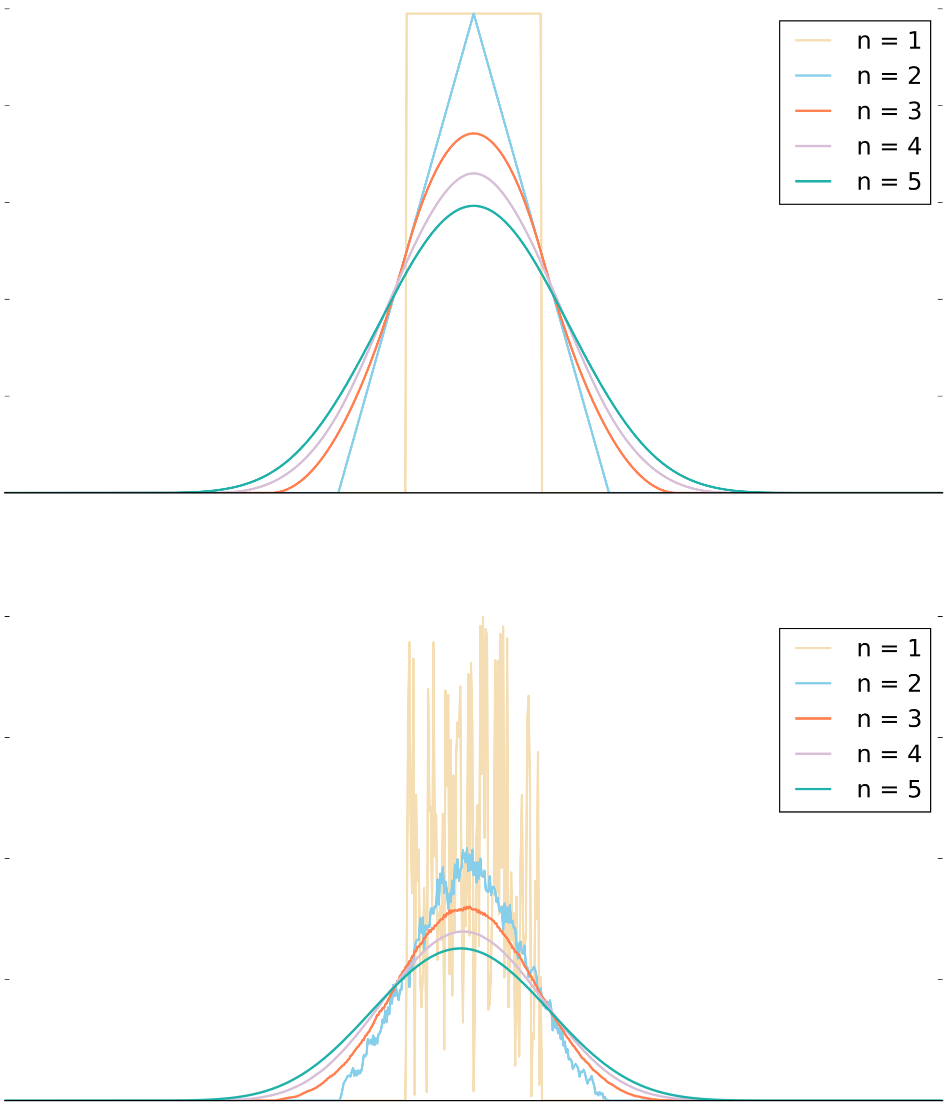


Figure 3: Result of convolving two different distributions with themselves several times. The shapes quickly become Gaussian-like.

If they are continuous

$$f_Z(z) = \int_{u=-\infty}^{\infty} f_X(u)\, f_Y(z-u)\; du \tag{26}$$

$$= (f_X * f_Y)(z). \tag{27}$$

The pmf (or pdf) of the sum is the convolution of the individual pmfs (or pdfs).

Corollary 1.21. *The pmf of the sample mean of an iid sequence is*

$$p_{\bar{X}}(x) = (p * p * \cdots * p)(x) \tag{28}$$

in the case of discrete sequences with pmf p. In the case of continuous sequences that have a pdf f, the pdf of the sample mean is

$$f_{\bar{X}}(x) = (f * f * \cdots * f)(x). \tag{29}$$

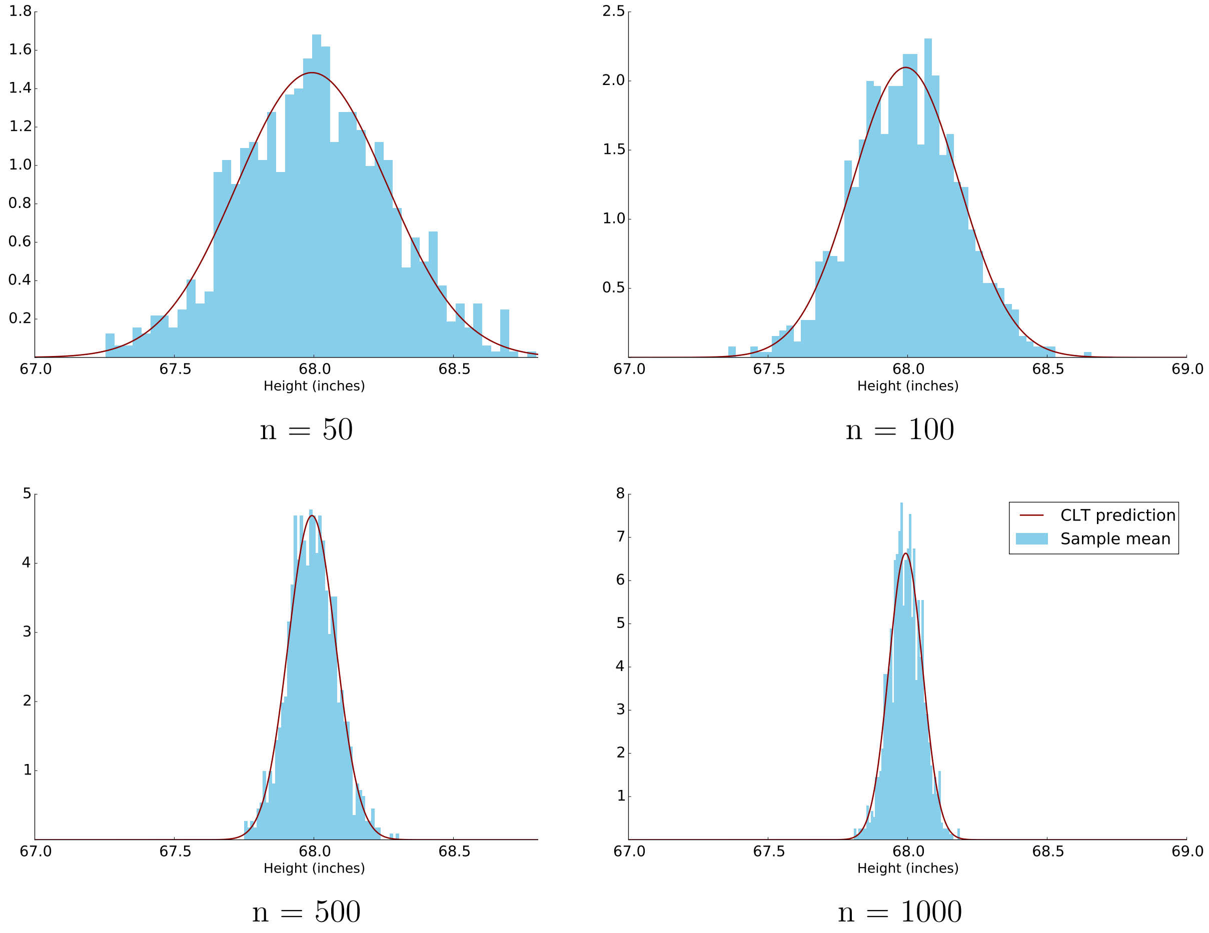
Repeated convolutions have a smoothing effect on the pmf or pdf of the sample mean as n increases. We show this numerically in Figure 3 for two very different distributions: a uniform distribution and a very irregular one. Both converge to Gaussian-like shapes after just 3 or 4 convolutions. The Central Limit Theorem makes this precise, establishing that the shape of the pmf or pdf becomes Gaussian asymptotically. This can be established rigorously by deriving the characteristic function of the sample mean and showing that it converges to that of a Gaussian; please see more advanced texts on probability theory for the details. □

Figure 4 shows the empirical pmf of the sample mean for data sampled at random from the dataset in Figure 1. The convergence to the Gaussian distribution predicted by the Central Limit Theorem is surprisingly fast.

The Central Limit Theorem not only allows to characterize estimators such as the sample mean very precisely. It also justifies the use of Gaussian distributions to model data that are the result of many different independent factors. For example, the height or weight of a population often has a Gaussian shape– as illustrated by Figure 1– because the height and weight of a person depends on many different factors that are roughly independent. Noise in many systems is well modeled as having a Gaussian distribution for the same reason.

1.5 Confidence intervals

The sample mean and variance are **point estimates** because they consist of a single value. In contrast, **confidence intervals** provide an interval estimate in which the parameter of interest lies with high probability.



n = 50

n = 100

n = 500

n = 1000

Figure 4: Empirical distribution of the sample mean for different values of n plotted together with the distribution predicted by the Central Limit Theorem.

Definition 1.22 (Confidence interval)**.** *A* $1-\alpha$ *confidence interval* $\mathcal{I}$ *for a parameter* γ *satisfies*

$$\mathrm{P}\left(\gamma \in \mathcal{I}\right) \geq 1-\alpha. \tag{30}$$

Note that the parameter γ is deterministic, whereas the interval is random. As a result, when we compute a confidence interval for a particular dataset it is incorrect to state that *the true parameter belongs to this particular interval with probability* $1-\alpha$ because the interval is a fixed **realization** and consequently the statement **does not include any random quantities!** The correct interpretation is that if we carry out many independent experiments and compute a 95% confidence interval for each of them, then the true parameter will lie in the interval 95% of the time (note that the experiments can be completely different).

Chebyshev's inequality allows to find an exact confidence interval for the mean of an iid sequence. The confidence interval is centered around the sample mean.

Theorem 1.23 (Confidence interval for the sample mean)**.** *If* $X_1, X_2, \ldots$ *is an iid sequence with bounded variance, such that* $\mathrm{E}\left(X_i\right) = \mu$ *and* $\sigma_{X_i} = \sigma^2$*, then for any* $0 < \alpha < 1$ $\left[\bar{X}_n - \frac{\sigma}{\sqrt{\alpha\, n}}, \bar{X}_n + \frac{\sigma}{\sqrt{\alpha\, n}}\right]$ *is a* $1-\alpha$ *confidence interval for* μ*, i.e.*

$$P\left(\mu \in \left[\bar{X}_n - \frac{\sigma}{\sqrt{\alpha\, n}}, \bar{X}_n + \frac{\sigma}{\sqrt{\alpha\, n}}\right]\right) \geq 1-\alpha. \tag{31}$$

Proof.

$$\begin{aligned} P\left(\mu \in \left[\bar{X}_n - \frac{\sigma}{\sqrt{\alpha\, n}}, \bar{X}_n + \frac{\sigma}{\sqrt{\alpha\, n}}\right]\right) &= 1 - P\left(\left|\bar{X}_n - \mu\right| > \frac{\sigma}{\sqrt{\alpha\, n}}\right) && (32)\\ &\geq 1 - \frac{\alpha\, n \mathrm{Var}\left(\bar{X}_n\right)}{\sigma^2} \quad \text{by Chebyshev's inequality} && (33)\\ &= 1-\alpha \quad \text{by Corollary 1.8.} && (34) \end{aligned}$$

□

The careful reader will notice that the confidence interval in Theorem 1.23 depends on the variance of the random variables in the sequence, which is unknown! The following corollary tackles this issue, showing that we can compute a confidence that only depends on an upper bound on the variance of the variables in the sequence.

Corollary 1.24 (Confidence interval for variables with bounded variance)**.** *If* $X_1, X_2, \ldots$ *is an iid sequence with bounded variance, such that* $\mathrm{E}\left(X_i\right) = \mu$ *and* $\sigma_{X_i} \leq b$*,then for any* $0 < \alpha < 1$ $\left[\bar{X}_n - \frac{\sigma}{\sqrt{\alpha\, n}}, \bar{X}_n + \frac{\sigma}{\sqrt{\alpha\, n}}\right]$ *is a* $1-\alpha$ *confidence interval for* μ*, i.e.*

$$P\left(\mu \in \left[\bar{X}_n - \frac{b}{\sqrt{\alpha\, n}}, \bar{X}_n + \frac{b}{\sqrt{\alpha\, n}}\right]\right) \geq 1-\alpha. \tag{35}$$

Proof. By Corollary 1.8 the variance of the sample mean is $\sigma^2/n \leq b^2/n$, so the argument in the proof of Theorem 1.23 follows through substituting σ by b. □

The Central Limit Theorem allows to characterize the distribution of the sample mean asymptotically. However, as we saw in Figures 3 and 4 it can be very accurate even for small values of n. This suggests leveraging the result to derive an *approximate* confidence interval that does not provide a rigorous guarantee like Theorem 1.23, but which can be much more precise.

Since the cdf of the Gaussian does not have a closed-form solution, we will write the asymptotic confidence interval in terms of the Q function, which is defined in Section D of the appendix as $Q(u) := P(U > u)$ for $u > 0$, where U is a Gaussian random variable with zero mean and unit variance. The following lemma, proved in Section E of the appendix, shows how to express probabilities concerning arbitrary Gaussian random variables using the Q function.

Lemma 1.25. *Let X be a Gaussian random variable with mean μ and variance σ. We have*

$$P(X > x) = Q\left(\frac{x-\mu}{\sigma}\right) \quad \text{for } x > \mu, \tag{36}$$

$$P(X < x) = Q\left(\frac{\mu-x}{\sigma}\right) \quad \text{for } x < \mu. \tag{37}$$

The following theorem provides an asymptotic confidence interval for the mean of a population. Now even the careless reader will remark that the interval depends on the unknown standard deviation σ. An option is to use the empirical variance to approximate the actual variance. In this case, the approximation will be significantly worse for small values of n.

Theorem 1.26 (Asymptotic confidence interval for the sample mean)**.** *If $X_1, X_2, \ldots$ is an iid sequence with bounded variance, such that $\mathrm{E}(X_i) = \mu$ and $\sigma_{X_i} = \sigma^2$, then for any $0 < \alpha < 1$ $\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}}Q^{-1}\left(\frac{\alpha}{2}\right), \bar{X}_n + \frac{\sigma}{\sqrt{n}}Q^{-1}\left(\frac{\alpha}{2}\right)\right]$ is an asymptotic $1-\alpha$ confidence interval for μ. More formally,*

$$P\left(\mu \in \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}}Q^{-1}\left(\frac{\alpha}{2}\right), \bar{X}_n + \frac{\sigma}{\sqrt{n}}Q^{-1}\left(\frac{\alpha}{2}\right)\right]\right) \approx 1-\alpha \tag{38}$$

as $n \to \infty$.

Proof. By the Central Limit Theorem, when $n \to \infty$ $\bar{X}_n$ is distributed as a Gaussian random variable with mean μ and variance σ^2. As a result

$$\begin{aligned}
& P\left(\mu \in \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right), \bar{X}_n + \frac{\sigma}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right)\right]\right) && (39)\\
&= 1 - P\left(\left|\bar{X}_n - \mu\right| > \frac{\sigma}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right)\right) && (40)\\
&= 1 - P\left(\bar{X}_n > \mu + \frac{\sigma}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right)\right) - P\left(\bar{X}_n < \mu - \frac{\sigma}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right)\right) && (41)\\
&\approx 1 - 2\,Q\left(Q^{-1}\left(\frac{\alpha}{2}\right)\right) \quad \text{by the CLT and Lemma 1.25} && (42)\\
&= 1 - \alpha. && (43)
\end{aligned}$$

□

Example 1.27 (Confidence intervals)**.** You want to estimate the average weight of the black bear population in Yosemite National Park. You manage to capture 300 bears, which you assume that is a uniform sample of the population. The sample mean is 200 lbs.

First you decide to compute a conservative 95% confidence interval. The maximum weight recorded for a black bear ever is 880 lbs. This means that if X_i is the weight of the ith bear

$$\begin{aligned}
\operatorname{Var}(X_i) &= \mathrm{E}\left(X_i^2\right) - \mathrm{E}^2(X_i) && (44)\\
&\leq \mathrm{E}\left(X_i^2\right) && (45)\\
&\leq 880^2 \quad \text{because } X_i \leq 880. && (46)
\end{aligned}$$

As a result, 880 is an upper bound for the standard deviation. Applying Corollary 1.24,

$$\left[\bar{X}_n - \frac{b}{\sqrt{\alpha\, n}}, \bar{X}_n + \frac{b}{\sqrt{\alpha\, n}}\right] = [-27.2, 427.2] \tag{47}$$

is a 95% confidence interval for the mean weight of the population.

Now you want to derive a more precise confidence interval, assuming that $n = 300$ is large enough for the sample variance, which is equal to $(100lbs)^2$, to be close to the true standard deviation and for the Central Limit theorem to provide a good approximation. To compute the interval we use the fact that $Q(1.95) \approx 0.025$ according to the Table in Section D of the appendix. By Theorem 1.26

$$\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right), \bar{X}_n + \frac{\sigma}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right)\right] \approx [188.8, 211.3] \tag{48}$$

is an asymptotic 95% confidence interval for the mean weight of the population (with the caveat that we are using the sample variance instead of the true variance).

A Proof of Theorem 1.7

The theorem is a direct consequence of the following identity.

$$\begin{aligned}
\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) &= \operatorname{E}\left(\sum_{i=1}^{n} X_i - \operatorname{E}\left(X_i\right)\right) && (49)\\
&= \sum_{i=1}^{n} \operatorname{E}\left(\left(X_i - \operatorname{E}\left(X_i\right)\right)^2\right) + \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} \operatorname{E}\left(\left(X_i - \operatorname{E}\left(X_i\right)\right)\left(X_j - \operatorname{E}\left(X_j\right)\right)\right) && (50)\\
&= \sum_{i=1}^{n} \operatorname{Var}\left(X_i\right) + \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} \operatorname{Cov}\left(X_i, X_j\right). && (51)
\end{aligned}$$

B Proof of Lemma 1.10

Let us define the mean square of the X_i

$$\mathrm{ms} := \sigma^2 + \mu^2 = \operatorname{E}\left(X_i^2\right). \tag{52}$$

By assumption if $i \neq j$ X_i and X_j are uncorrelated so

$$\operatorname{E}\left(X_i X_j\right) = \operatorname{E}\left(X_i\right)\operatorname{E}\left(X_j\right) = \mu^2. \tag{53}$$

Expanding the definition of the sample mean,

$$\begin{aligned}
\left(X_i - \bar{X}\right)^2 &= \left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)^2 && (54)\\
&= X_i^2 + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} X_j X_k - \frac{2}{n}\sum_{j=1}^{n} X_i X_j && (55)
\end{aligned}$$

As a result, we have

$$\mathrm{E}\left(S^2\right) = \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{E}\left(X_i^2\right) + \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{E}\left(X_j^2\right) + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\mathrm{E}\left(X_j X_k\right) \tag{56}$$

$$- \frac{2}{n}\mathrm{E}\left(X_i^2\right) - \frac{2}{n}\sum_{\substack{j=1\\j\neq i}}^{n}\mathrm{E}\left(X_i X_j\right) \tag{57}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} \mathrm{ms} + \frac{n\,\mathrm{ms}}{n^2} + \frac{n\left(n-1\right)\mu^2}{n^2} - \frac{2\,\mathrm{ms}}{n} - \frac{2\left(n-1\right)\mu^2}{n} \tag{58}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\frac{n-1}{n}\left(\mathrm{ms} - \mu^2\right) \tag{59}$$

$$= \sigma^2. \tag{60}$$

C Proof of Lemma 1.20

In the discrete case,

$$p_Z\left(z\right) = \mathrm{P}\left(Z = z\right) \tag{61}$$

$$= \mathrm{P}\left(X + Y \leq z\right) \tag{62}$$

$$= \sum_{u=-\infty}^{\infty}\mathrm{P}\left(X = u, Y = z - u\right) \quad \text{disjoint events} \tag{63}$$

$$= \sum_{u=-\infty}^{\infty} p_X\left(u\right) p_Y\left(z-u\right). \tag{64}$$

In the continuous case,

$$F_Z\left(z\right) = \mathrm{P}\left(Z \leq z\right) \tag{65}$$

$$= \mathrm{P}\left(\cup_{u=-\infty}^{\infty}\left\{X = u, Y = z - u\right\}\right) \tag{66}$$

$$= \int_{x=\infty}^{\infty}\int_{y=\infty}^{z-x} f_{X,Y}\left(x,y\right)\,\mathrm{d}x\,\mathrm{d}y \tag{67}$$

$$= \int_{x=\infty}^{\infty}\int_{y=\infty}^{z-x} f_X\left(x\right) f_Y\left(y\right)\,\mathrm{d}x\,\mathrm{d}y \tag{68}$$

$$= \int_{x=\infty}^{\infty} f_X\left(x\right) F_Y\left(z-x\right)\,\mathrm{d}x. \tag{69}$$

The result follows by differentiating with respect to z.

D The Q function

The $Q(\cdot)$ function is the area beneath the righthand tail of the pdf of a Gaussian random variable with zero mean and unit variance:

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\, dx \ . \tag{70}$$

The following table lists values of $Q(x)$ for $0 \le x \le 4$.

x	$Q(x)$	x	$Q(x)$
0.0	0.50000	2.0	2.2750×10^{-2}
0.1	0.46017	2.1	1.7864×10^{-2}
0.2	0.42074	2.2	1.3903×10^{-2}
0.3	0.38209	2.3	1.0724×10^{-2}
0.4	0.34458	2.4	8.1975×10^{-3}
0.5	0.30854	2.5	6.2097×10^{-3}
0.6	0.27425	2.6	4.6612×10^{-3}
0.7	0.24196	2.7	3.4670×10^{-3}
0.8	0.21186	2.8	2.5551×10^{-3}
0.9	0.18406	2.9	1.8658×10^{-3}
1.0	0.15866	3.0	1.3499×10^{-3}
1.1	0.13567	3.1	9.6760×10^{-4}
1.2	0.11507	3.2	6.8714×10^{-4}
1.3	0.09680	3.3	4.8342×10^{-4}
1.4	0.08076	3.4	3.3693×10^{-4}
1.5	0.06681	3.5	2.3263×10^{-4}
1.6	0.05480	3.6	1.5911×10^{-4}
1.7	0.04457	3.7	1.0780×10^{-4}
1.8	0.03593	3.8	7.2348×10^{-5}
1.9	0.02872	3.9	4.8096×10^{-5}
2.0	0.02275	4.0	3.1671×10^{-5}

For $x > 4$, the function can be approximated by

$$Q(x) \approx \frac{1}{x\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \ .$$

E Proof of Lemma 1.25

By Theorem 2.1 $U := \frac{X-\mu}{\sigma}$ is a Gaussian random variable with zero mean and unit variance. For $x > \mu$

$$\begin{aligned} P(X > x) &= P(\sigma U + \mu > x) && (71)\\ &= P\left(U > \frac{x-\mu}{\sigma}\right) && (72)\\ &= Q\left(\frac{x-\mu}{\sigma}\right). && (73) \end{aligned}$$

For $x < \mu$

$$\begin{aligned} P(X < x) &= P(\sigma U + \mu < x) && (74)\\ &= P\left(U < \frac{x-\mu}{\sigma}\right) && (75)\\ &= P\left(U > \frac{\mu-x}{\sigma}\right) \quad \text{by symmetry of the Gaussian pdf} && (76)\\ &= Q\left(\frac{\mu-x}{\sigma}\right). && (77) \end{aligned}$$