15 - Estimating moments

- The central limit theorem
- Example: sums of IID random variables
- Estimating the mean of a random vector
- Example: estimating the mean of a Gaussians
- The Chebyshev bound and the χ^2 test
- The Chebyshev bound for Gaussians
- Example: estimating frequencies
- Estimating the covariance for arbitrary distributions
- The sample covariance
- The sample covariance is unbiased
- Estimating the mean when the covariance is unknown
- The student's t-distribution
- Confidence intervals with unknown covariance

The central limit theorem

Suppose $x_1, x_2, \ldots$ are IID random variables, each with mean μ, variance σ^2. Define the *sample mean* s_n and *normalized sample mean* z_n

$$s_n = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad z_n = \frac{\sqrt{n}}{\sigma}(s_n - \mu)$$

Notice that

- Both s_n and z_n are *random variables*
- $\mathbf{E}\, s_n = \mu$ and $\mathbf{cov}(s_n) = \dfrac{\sigma^2}{n}$
- $\mathbf{E}\, z_n = 0$ and $\mathbf{cov}(z_n) = 1$

The central limit theorem

The surprising fact is that s_n and z_n are *asymptotically Gaussian*; that is

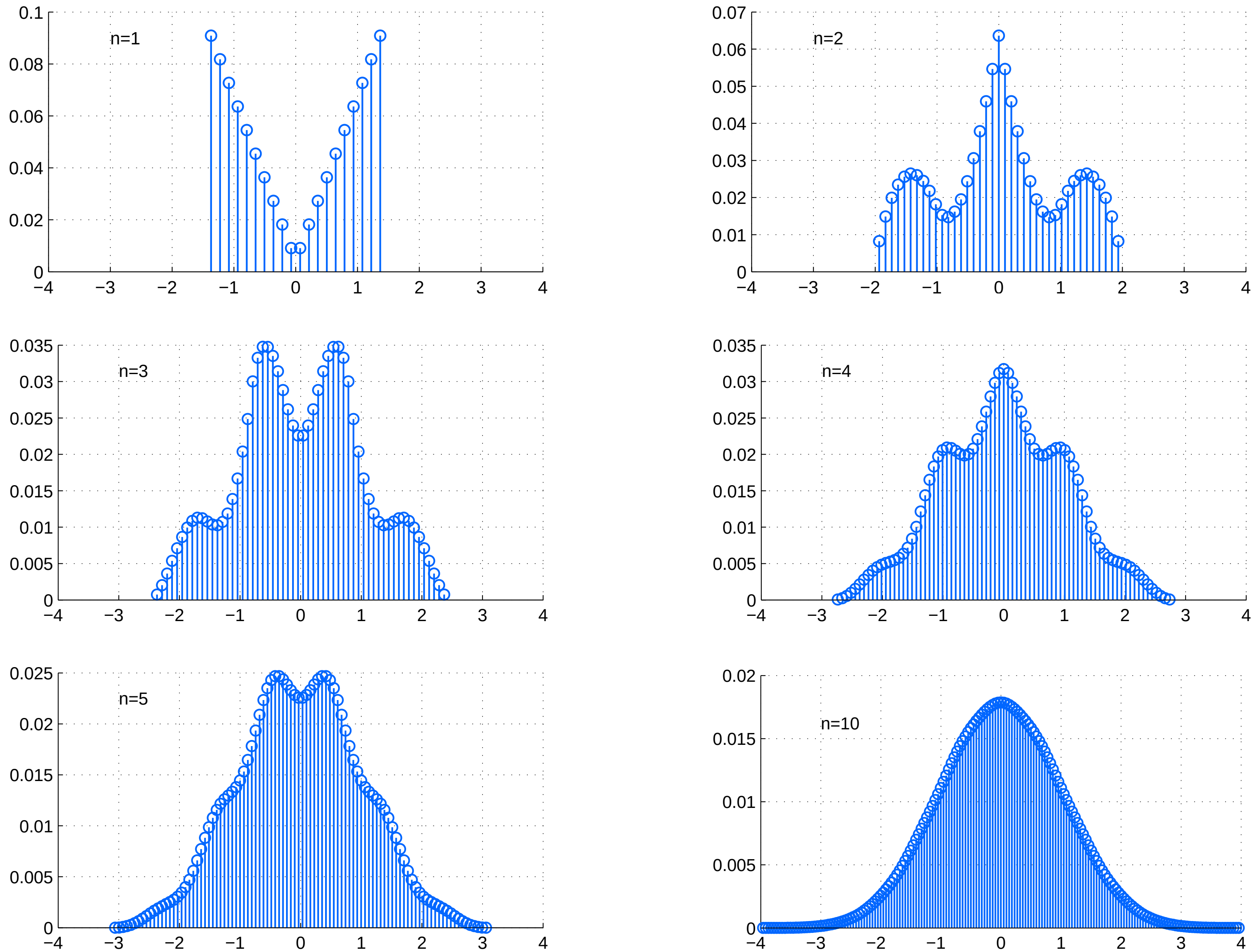
$$\boxed{\lim_{n\to\infty} \mathbf{Prob}(z_n \leq a) = F_{\mathcal{N}}(a)}$$

Here $F_{\mathcal{N}}$ is the cdf of a Gaussian with mean 0 and covariance 1.

$$F_{\mathcal{N}}(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx$$

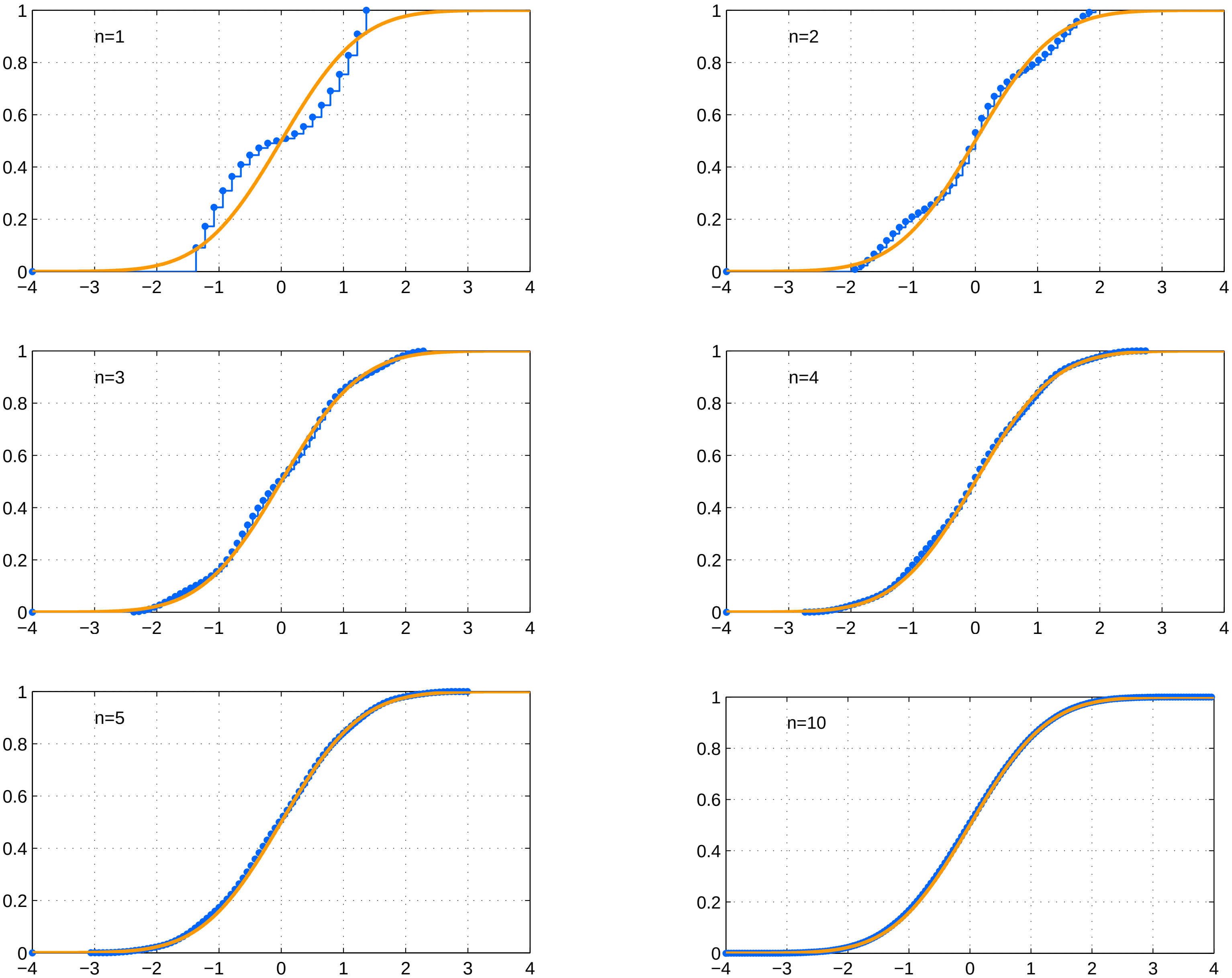
Sums of IID random variables

The pmf of z_n for various values of n; note the graph tends to a 'Gaussian shape', although the pmf is *discrete*



Sums of IID random variables

The cdf of z_n for various values of n; orange curve is $F_{\mathcal{N}}$



Estimating the mean of a random vector

$x_1, x_2, \ldots,$ are IID random vectors $x_i : \Omega \to \mathbb{R}^m$ with mean μ and covariance Σ.

We can *estimate* the mean μ, using the *sample mean* s_n

$$s_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

This has the properties that

- s_n is *unbiased*, i.e., its expected value is correct

$$\mathbf{E}\, s_n = \mu$$

- s_n is *consistent*, i.e, as the number of measurements becomes large, the probability of an error of ε shrinks to zero

$$\text{for any } \varepsilon > 0 \qquad \lim_{n \to \infty} \mathbf{Prob}\big(|s_n - \mu| \geq \varepsilon\big) = 0$$

Example: estimating the mean of a Gaussian

The covariance of the sample mean is

$$\mathbf{cov}(s_n) = \frac{\Sigma}{n}$$

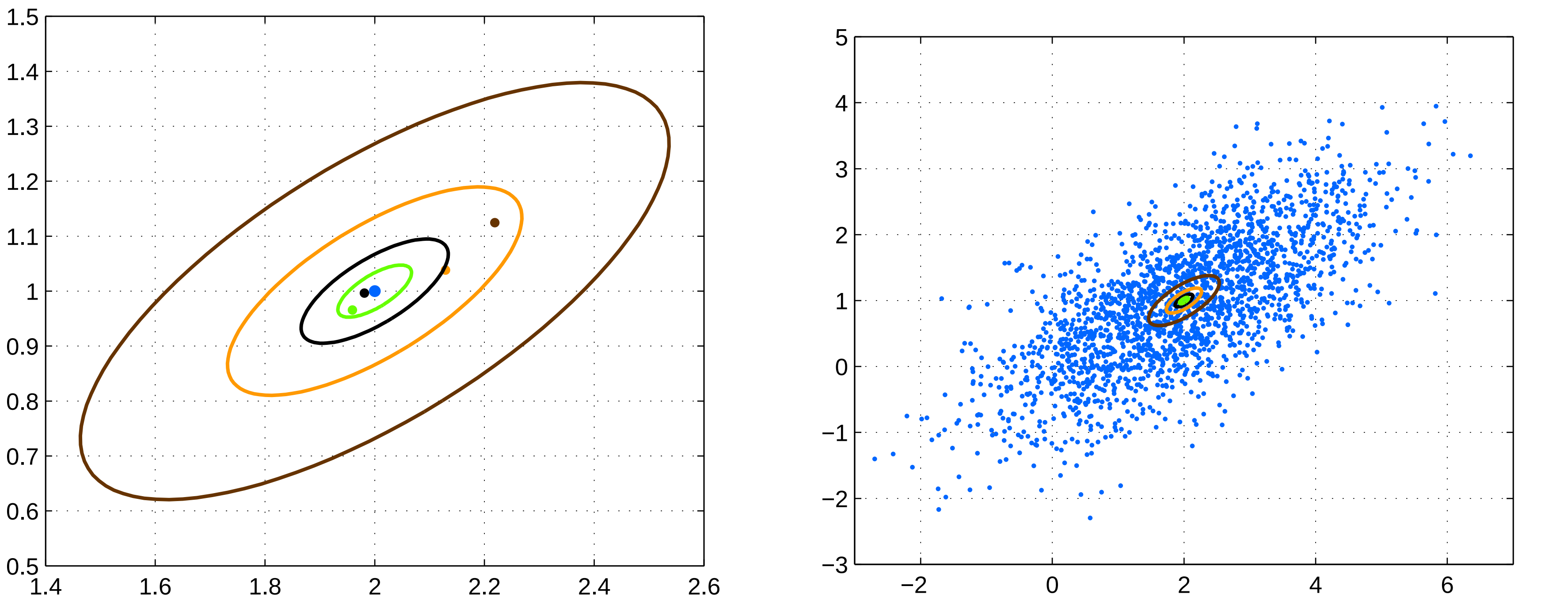
If the x_i are Gaussian then we have the 90% confidence ellipsoids

$$\mathbf{Prob}\left(s_n^T \Sigma^{-1} s_n \leq \frac{1}{n} F_{\chi_m^2}^{-1}(0.9)\right) = 0.9$$

Estimating the mean of a Gaussian

suppose $\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}, \mu = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$

2048 data points, along with the ellipsoids for $n = 32, 128, 512, 2048$.



Estimating the mean for arbitrary distributions

Suppose $x_1, x_2, \ldots$ are IID random variables, each with mean μ, variance σ^2.

The Chebyshev inequality gives a confidence bound for the sample mean

$$\mathbf{Prob}\big(|s_n - \mu| \leq \varepsilon\big) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}$$

because

$$\mathbf{cov}(s_n) = \frac{\mathbf{cov}(x_i)}{n}$$

- Above is true irrespective of the pdf on x_i
- However s_n tends to a Gaussian as n becomes large
- So instead of using the Chebyshev bound, we can use the χ^2 confidence bound
- often called *the χ^2 test*

The Chebyshev bound and the χ^2 test

Suppose $x_i : \Omega \to \mathbb{R}$ are IID random variables, each with mean μ and covariance σ^2.

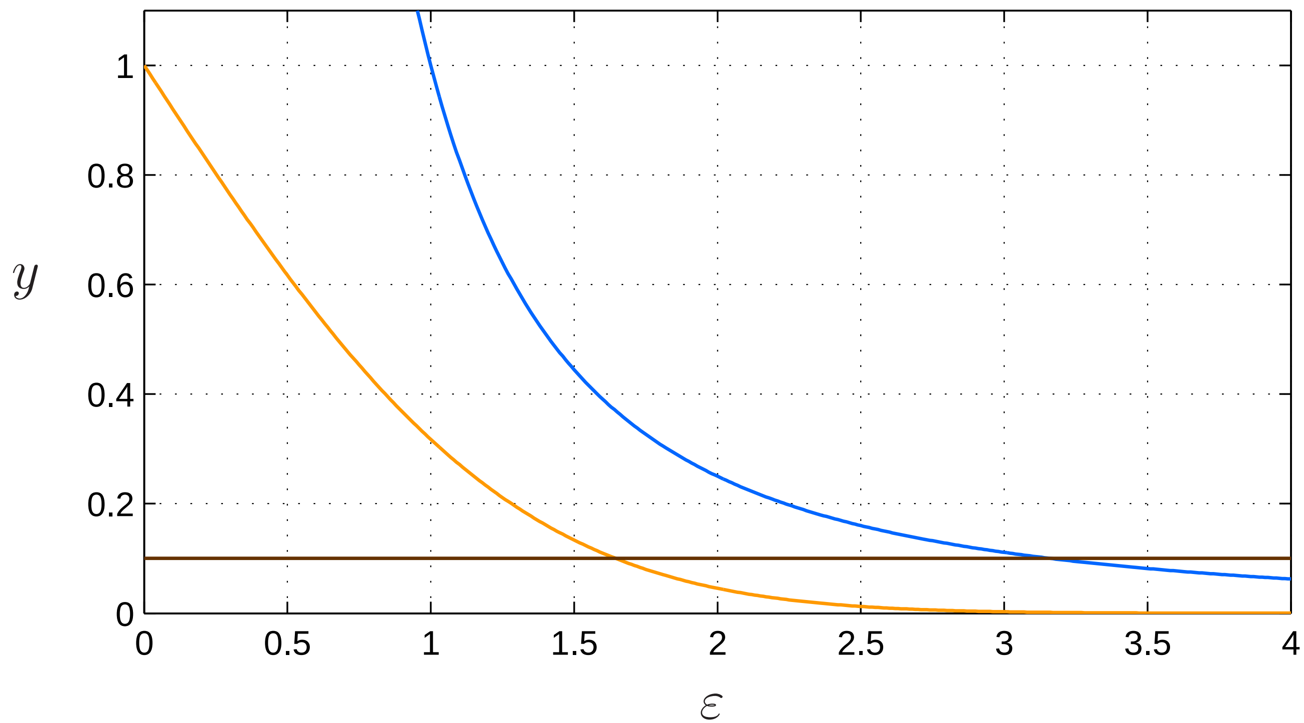
The Chebyshev inequality gives

$$\mathbf{Prob}\big(|s_n - \mu| \leq \varepsilon\big) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}$$

But we know for large n that s_n is close to Gaussian, so

$$\mathbf{Prob}\big(|s_n - \mu| \leq \varepsilon\big) \approx F_{\chi_1^2}\left(\frac{n\varepsilon^2}{\sigma^2}\right)$$

The Chebyshev bound for Gaussians



- $x \sim \mathcal{N}(0,1)$
- Orange curve is $y = 1 - F_{\chi_1^2}(\varepsilon^2)$, that is $y = \mathbf{Prob}(|x - \mathbf{E}\, x| \geq \varepsilon)$
- Blue curve is the Chebyshev bound $y = \varepsilon^{-2}$
- 90% confidence with χ^2 for $\varepsilon \approx 1.65$, with Chebyshev is $\varepsilon \approx 3.16$

The Chebyshev bound and the χ^2 test

- For probability p, with Chebyshev the confidence interval half-width is

$$\varepsilon_{\text{cheby}} = \frac{\sigma}{\sqrt{n(1-p)}}$$

- With χ^2 the confidence interval half-width is

$$\varepsilon_{\text{chi}} = \frac{\sigma\sqrt{F_{\chi_1^2}^{-1}(p)}}{\sqrt{n}}$$

Although the χ^2 bound is tighter, both scale as $\dfrac{1}{\sqrt{n}}$

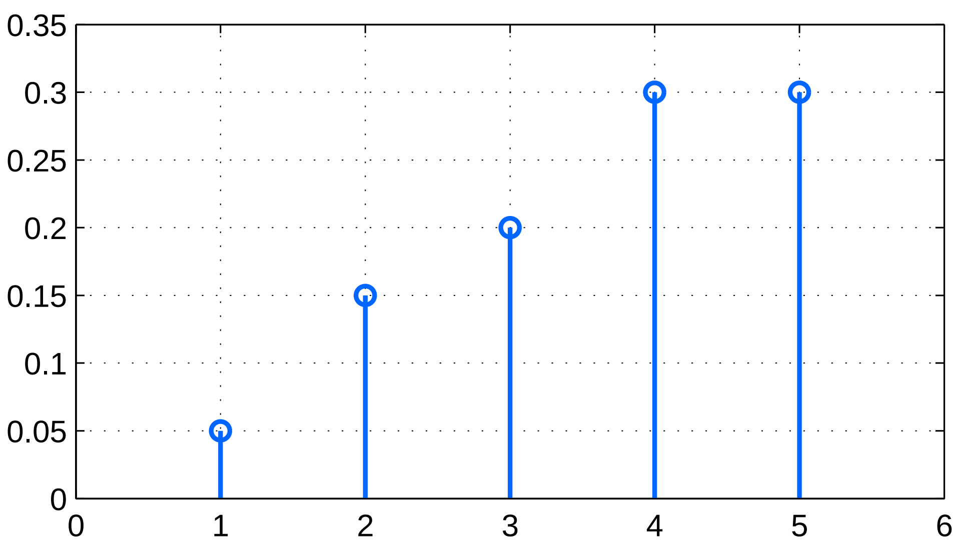
For both bounds, we need to know the covariance; we'll fix this, using the t-test ...

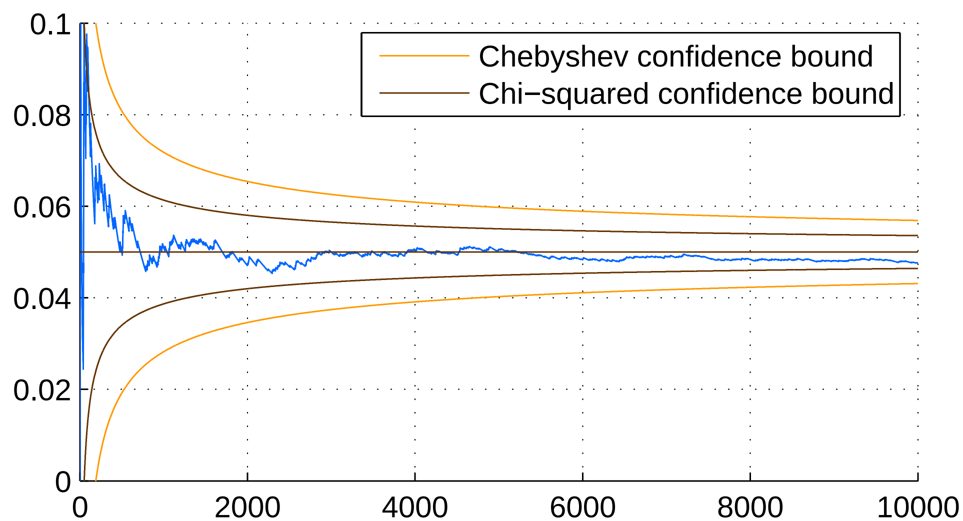
Example: estimating frequencies

We can use this for example when estimating the pmf of a discrete random variable

- the *frequency* $s_n = \frac{1}{n}\sum_{i=1}^{n} I_j$ is a sum of IID *indicator functions* I_j, each of which is a Bernoulli random variable
- hence the frequencies are approximately Gaussian for large n

The confidence bounds given by Chebyshev and χ^2 are shown below.





Estimating covariance for arbitrary distributions

How do we estimate Σ? The answer depends on whether we know μ or not.

If we know μ, let

$$T_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$$

- T_n is a *unbiased estimate* of Σ, i.e.,

$$\begin{aligned} \mathbf{E}\, T_n &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\big((x_i - \mu)(x_i - \mu)^T\big) \\ &= \frac{1}{n}n\Sigma \\ &= \Sigma \end{aligned}$$

- T_n is also *consistent*, by the law of large numbers
- For confidence bounds, we would need the distribution.

The sample covariance

The *sample covariance* Q_n is

$$\boxed{Q_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - s_n)(x_i - s_n)^T}$$

- When we don't know μ, we use the sample mean s_n in place of the mean μ
- But, the factor in front of the sum is $\dfrac{1}{n-1}$ not $\dfrac{1}{n}$.
- With this choice, the estimate is unbiased.

Proof that the sample covariance is unbiased

we'd like to find $\mathbf{E}\, Q_n$; we have

$$
\begin{aligned}
Q_n &= \frac{1}{n-1}\sum_{i=1}^n (x_i - s_n)(x_i - s_n)^T \\
&= \frac{1}{n-1}\sum_{i=1}^n \Big((x_i-\mu)(x_i-\mu)^T - \mu\mu^T + \mu x_i^T + x_i\mu^T - s_n x_i^T - x_i s_n^T + s_n s_n^T \Big) \\
&= \frac{1}{n-1}\sum_{i=1}^n \Big((x_i-\mu)(x_i-\mu)^T \Big) \; - \; \frac{n}{n-1}(s_n-\mu)(s_n-\mu)^T
\end{aligned}
$$

- because $s_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^n x_i$
- similar to $\mathbf{E}\big(\|x\|^2\big) = \mathbf{trace}\big(\mathbf{cov}(x)\big) + \|\mathbf{E}\, x\|^2$

Proof that the sample covariance is unbiased

the expectation of the second term is

$$\begin{aligned}\mathbf{E}\big((s_n-\mu)(s_n-\mu)^T\big) &= \mathbf{cov}(s_n)\\ &= \frac{\Sigma}{n}\end{aligned}$$

so we have

$$\begin{aligned}\mathbf{E}\,Q_n &= \frac{1}{n-1}\sum_{i=1}^n \mathbf{E}\big((x_i-\mu)(x_i-\mu)^T\big) \;-\; \frac{1}{n-1}\Sigma\\ &= \frac{1}{n-1}\sum_{i=1}^n \Sigma \;-\; \frac{1}{n-1}\Sigma\\ &= \Sigma\end{aligned}$$

- Hence the sample covariance Q_n is an *unbiased estimate* of the covariance Σ
- it can also be shown that Q_n is *consistent*

Estimating the mean when the covariance is unknown

We saw earlier that one can construct confidence intervals as follows

- The sample means s_n are approximately Gaussian
- Hence we have the confidence bounds

$$\mathbf{Prob}\left(\frac{|s_n - \mu|\sqrt{n}}{\sigma} \leq \varepsilon\right) \approx F_{\chi_1^2}(\varepsilon^2)$$

- We can use these to *estimate the mean* μ

But to do this, we need to know the *covariance* σ^2

What do we do when we do not know σ^2?

The student's t-distribution

scalar random variable $x \in \mathbb{R}$ is called *t-distributed with n degrees of freedom* if

$$f_{t_n}(x) = C_n \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

where C_n is the normalizing constant

$$C_n = \frac{\Gamma\big((n+1)/2\big)}{\sqrt{n\pi}\,\Gamma(n/2)}$$

(matlab function `tpdf`, `tcdf` and `tinv`)

named after William Gosset, who was a chemist for Guinness brewery in Dublin from 1899 to 1935. Guinness would not let him publish under his own name

he invented the t-distribution for quality control in brewing

Confidence intervals with unknown covariance

When we do not know σ^2, we use instead the *sample covariance* Q_n

if $x_1, \dots, x_n$ are scalar Gaussian random variables $x_i \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\frac{(s_n - \mu)\sqrt{n}}{\sqrt{Q_n}}$$

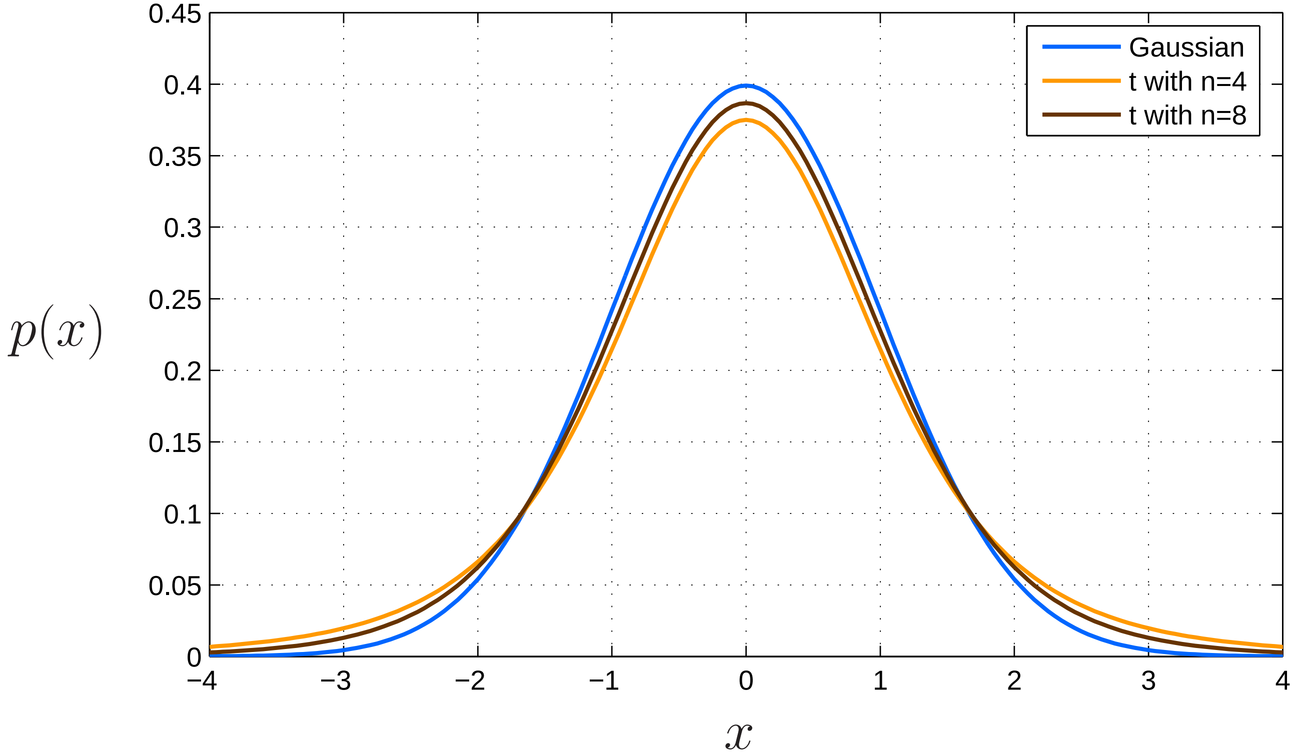
has a t-distribution with $n-1$ degrees of freedom, so

$$\begin{aligned}\textbf{Prob}\left(-z \leq \frac{(s_n - \mu)\sqrt{n}}{\sqrt{Q_n}} \leq z\right) &= \int_{-z}^{z} f_{t_{n-1}}(x)\,dx \\ &= 1 - 2F_{t_{n-1}}(-z)\end{aligned}$$

the confidence interval width grows *approximately* as $\dfrac{1}{\sqrt{n}}$

Comparison of known and unknown covariance

The pdfs for $\mathcal{N}(0,1)$ and the t-distribution are below



- the t pdf is slightly wider than the Gaussian
- because we have less information about the mean when the variance is unknown
- as the no. of measurements n becomes large, the t-pdf tends to the Gaussian pdf

Distributions and densities in Matlab

Matlab has useful functions in the statistics toolbox:

`chi2pdf`	Chi square density
`normpdf`	Gaussian density
`tpdf`	t-density
`chi2cdf`	Chi square cdf
`normcdf`	Gaussian cdf
`tcdf`	t-cdf
`chi2inv`	Chi square inverse cdf
`norminv`	Gaussian inverse cdf
`tinv`	t-inverse cdf

as well as `gamma` and `erf`