Lecture Notes: Chapter 4

20/21

4 Sampling

4.1. Sample Mean and Central Limit Theorem

Suppose we take a series of measurements from some population (e.g., height, duration, etc.) Suppose the quantity we are measuring is distributed with mean μ and variance σ^2.

The sequence of measurements can be modeled as a sequence of RVs $X_1, X_2, \ldots, X_n$ that are i.i.d.

The n-th sample mean is the RV

$$\overline{X_n} := \frac{\sum_{i=1}^{n} X_i}{n}$$

We know that

$$E[\bar{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\cdot n\cdot\mu = \mu$$

$$Var(\bar{X}_n) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\cdot n\cdot\sigma^2 = \frac{1}{n}\sigma^2$$

That is,

- the mean stays the same
- the standard deviation approaches 0

This is the reason behind the weak law of large numbers:

$$P\left[\,|\bar{X}_n - \mu| > \varepsilon\,\right] \longrightarrow 0 \qquad (n \to \infty)$$

for all possible bounds $\varepsilon > 0$.

What is the Shape of the Distribution of $\overline{X}_n$?

Problem: $\overline{X}_n$ is squeezed by the division by n.

Consider instead

$$Y_i := \frac{X_i - \mu}{\sigma}.$$

Then $E[Y_i] = 0$, $Var(Y_i) = Var\left(\frac{X_i - \mu}{\sigma}\right) = \frac{1}{\sigma^2} Var(X_i) = 1.$

The X_i are i.i.d, so also the Y_i are i.i.d.

Let $$U_n := \frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}} = \sqrt{n} \cdot \overline{Y}_n = \sqrt{n}\, \frac{\overline{X}_n - \mu}{\sigma}$$

Then $$E[U_n] = \sqrt{n} \cdot E[\overline{Y}_n] = \sqrt{n} \cdot 0 = 0$$

$$Var(U_n) = Var(\sqrt{n}\, \overline{Y}_n) = n\, Var(\overline{Y}_n) = n \cdot \frac{1}{n} \cdot 1 = 1$$

The Central Limit Theorem (CLT)

The CLT says that the distributions of the U_n, (i.e., the cdfs) converge towards the cdf of the standard normal.

Theorem (Lindeberg-Lévy) [Central Limit Theorem]

Let X_i be i.i.d. RVs with mean μ and variance σ^2 and let

- $U_n := \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)$
- F_n be the cdf of U_n (i.e., $F_n(x) = P[U_n \leq x]$)
- Φ be the cdf of $N(0,1)$.

Then

$$\lim_{n\to\infty} F_n(x) = \Phi(x) \quad \text{f.a. } x \in \mathbb{R}$$

Convergence in Distribution

This kind of convergence is called "convergence in distribution", which is the weakest kind of convergence among RVs.

For instance, the Weak Law of Large Numbers says that $\overline{X}_n \longrightarrow \mu$ "in probability", which implies convergence in distribution.

The CLT says, $F_n(x) \longrightarrow \Phi(x)$, but this may be fast for some x and slow for others.

In practice, convergence is faster for x close to 0, that is, close to the mean, and slow if $|x|$ is large, i.e., far away from the mean.

Interpretation and Application of the CLT

Let X_i be i.i.d. RVs with mean μ and variance σ^2.

Let $S_n := \sum_{i=1}^{n} X_i$ be the sum of the X_i.

The CLT says that for large n the normalized sum

$$\frac{1}{\sqrt{n}\,\sigma}(S_n - n\mu)$$

has approximately a standard normal distribution.

From that one can conclude that

$$S_n \sim N(n\mu, n\sigma^2) \text{ approximately},$$

where the approximation is best around the mean $n\mu$.

Probabilities of the values of S_n can then be approximated by probabilities of a normally distributed RV.

Example 64: An insurance company has 25,000 policy holders.

Considering the yearly claim of a policy holder as a RV, the company has observed that

- the mean of the claims is $\mu = € 320$
- the standard deviation is $\sigma = € 540$

What is the probability that the total yearly claim is $> € 8.3$ Mio ?

Let C_i be the yearly claim of policy holder i, and

$S_n = \sum_{i=1}^{n} C_i$ be the yearly sum of claims, $n = 25{,}000$.

$\overline{C}_n = \frac{1}{n} S_n$ be the average of the claims.

We want to know $P[S_n > s]$, where $s = 8.3$ Mio.

From the CLT, we conclude that

$$S_n \sim N(n\cdot\mu, n\sigma^2) \text{ approx.}$$

Hence

$$P[S_n > s] = P\left[\frac{S_n - n\mu}{\sqrt{n}\,\sigma} > \frac{s-n\mu}{\sqrt{n}\,\sigma}\right]$$

$$\approx P\left[Z > \frac{s-n\mu}{\sqrt{n}\,\sigma}\right] = 1 - \Phi\left(\frac{s-n\mu}{\sqrt{n}\,\sigma}\right)$$

Now:

$$n\mu = 25{,}000 \times 320$$
$$= 8 \times 10^6$$

$$\sqrt{n}\,\sigma = \sqrt{25{,}000} \times 540$$
$$= \sqrt{2.5} \times 5.4 \times 10^2 \times 10^2$$

$$s - n\mu = 8.3 \times 10^6 - 8 \times 10^6$$
$$= 3 \times 10^5$$

$$\frac{s-n\mu}{\sqrt{n}\,\sigma} = \frac{3}{\sqrt{2.5} \times 5.4} \, \frac{10^5}{10^4}$$
$$= 0.351 \times 10 = 3.51$$

Thus $P[S_n > s] = 1 - \Phi(3.51) = 1 - 0.9998 = 0.0002$

If we have access to R, we observe

$$S_n \sim \mathcal{N}(n \cdot \mu, n\sigma^2)$$

and we want to know

$$P[S_n > s] = 1 - P[S_n \leq s],$$

which is computed by the cell

$$1 - \text{dnorm}(s, n \cdot \mu, \sqrt{n}\,\sigma)$$

Recall that R requires the standard deviation as argument

Normal and Binomial Distribution

Corollary: Let X_i be independent Bernoulli(p) RVs. Then

$$\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{n} \cdot \sqrt{p \cdot (1-p)}} \longrightarrow N(0,1) \quad \text{in distribution}$$

in distribution.

Rules of Thumb: A Binomial(n,p) distribution is close to

- $N(np, np(1-p))$ if both $np > 5$, and $n(1-p) > 5$
- Poisson(np) if $np < 5$ or $n(1-p) < 5$, and $n > 20$

Example 65:

An airplane fits 150 passengers.

On a busy route, only 30 % of the people that buy a ticket take the plane.

If the airline sells 450 tickets per flight, what is the probability that the plane is overbooked?

The number of passenger $\mathcal{P}$ taking the plane is a binomial RV with mean $n \cdot p$ and variance $n \cdot p(1-p)$ where

$$n = 450 \, , \quad p = 0.3 .$$

Let $s = 150$ be the number of seats available.

The plane is overbooked if

$$\mathcal{P} > 150 .$$

We can approximate $\mathcal{P}$ by a RV $\mathcal{X} \sim \mathcal{N}(np, np(1-p))$. Then

$$P[\mathcal{P} > s] = P[\mathcal{X} > s + 0.5]$$

adjustment when translating a discrete *) into a continuous problem

$$= P\left[\frac{\mathcal{X} - np}{\sqrt{n}\sqrt{p(1-p)}} > \frac{s+0.5-np}{\sqrt{n}\sqrt{p(1-p)}}\right] = 1 - \Phi\left(\frac{s+0.5-np}{\sqrt{n}\sqrt{p(1-p)}}\right)$$

$$= 1 - \Phi(1.59) = 1 - 0.944 = 0.056 = 5.6\%$$

Alternatively, with R we could have called

$$1 - \mathrm{dnorm}(s+0.5, np, \sqrt{n\,p(1-p)})$$

*) called continuity correction

Example 69: Opinion Polling

Suppose that 40 % of the population support a certain political candidate.

Given a random sample of 150 individuals, find

1.) the expected value and variance of the number of sampled individuals that favour the candidate.

2.) the probability that more than half the sample favour the candidate.

Example 69: Opinion Polling

Suppose that 40 % of the population support a certain political candidate.
Given a random sample of 150 individuals, find

1.) the expected value and variance of the number of sampled individuals that favour the candidate.

2.) the probability that more than half the sample favour the candidate.

Let X_i be the answer of the i-th person, "yes" meaning 1, and "no" meaning 0.

$\Rightarrow$ $X_i \sim \text{Bernoulli}(p)$ with $p = 0.4$

Let $Y := \sum_{i=1}^{n} X_i \Rightarrow Y \sim \text{Binom}(n, p)$, with $n = 150$

$$\Rightarrow \quad E[Y] = n \cdot p = 150 \times 0.4 = 60$$

$$Var(Y) = n \cdot p \cdot (1-p) = 150 \times 0.4 \times 0.6 = 36$$

Check the rule of thumb:

$$n \times p = 60 > 5, \qquad n \times (1-p) = 90 > 5$$

$\Rightarrow$ Approximate Y by $N(60, 36)$.

We want

$$P[Y > 75]$$

How can we compute this?

1) Use the Binomial:

Let ψ be the cdf of $\text{Binom}(150, 0.4)$.

R delivers

$$P[Y > 75] = 1 - P[Y \leq 75] = 1 - \psi(75)$$

$$= 0.005225$$

2) Approximate Y by a $Y' \sim N(60, 36)$

$$P[Y > 75] \approx P[Y' > 75.5]^{*} = 1 - \Phi_{60,36}(75.5)$$

$$= 0.004892 \qquad \text{(with R)}$$

$$1 - \Phi\left(\frac{75.5 - 60}{6}\right)$$

* continuity correction

3) Approximation and Lookup in Z-Table

Transform Y' to $Z \sim \mathcal{N}(0,1)$:

$$P[Y' > 75.5] = P\left[\frac{Y'-60}{6} > \frac{75.5-60}{6}\right]$$

$$\approx P\left[Z > \frac{75.5-60}{6}\right] = P\left[Z > \frac{15.5}{6}\right]$$

$$= P[Z > 2.583] = 1 - \Phi(2.583)$$

$$\approx 1 - 0.9951 = 0.0049$$

How Many Measurements are Needed?

We can use the CLT to determine the number of measurements needed for a required accuracy if we know the variance of the distribution of measurements.

Example 66: We want to measure the distance to a star with

- accuracy $a = 1$ (i.e., with absolute error $\le \frac{a}{2} = 0.5$) and
- certainty $\gamma = 95\%$.

The variance of the measurements is $\sigma^2 = 2^2$.

Let d be the exact distance and X_i be the measurements.

The sample mean $\overline{X}_n$ is close to a normal with

$$\mu_n = \mu \quad \text{and} \quad \sigma_n^2 = \frac{\sigma^2}{n}.$$

Then

$$\frac{\overline{X}_n - \mu_n}{\sigma_n} = \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} \sim N(0,1) \text{ approximately.}$$

We want n such that

$$P\left[-\frac{a}{2} < \bar{X}_n - \mu < \frac{a}{2}\right] \leq \gamma$$

That is

$$\gamma \leq P\left[-\frac{\sqrt{n}}{\sigma}\frac{a}{2} < \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) < \frac{\sqrt{n}}{\sigma}\frac{a}{2}\right]$$

$$\approx P\left[-\frac{\sqrt{n}}{\sigma}\frac{a}{2} < Z < \frac{\sqrt{n}}{\sigma}\frac{a}{2}\right]$$

$$= 1 - 2\left(1 - \Phi\left(\sqrt{n}\frac{a}{2\sigma}\right)\right) = 2\cdot\Phi\left(\sqrt{n}\frac{a}{2\sigma}\right) - 1,$$

hence

$$\Phi\left(\sqrt{n}\frac{a}{2\sigma}\right) \geq \frac{1+\gamma}{2}$$

$$\iff \sqrt{n}\frac{a}{2\sigma} \geq \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$$

$$\iff \sqrt{n} \geq \frac{2\sigma}{a}\Phi^{-1}\left(\frac{1+\gamma}{2}\right)$$

This is an example where we need the inverse of the cdf to reason backward from a probability to an argument.

We need an n such that

$$\sqrt{n} \geq \frac{2\sigma}{a} \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$$

with

$$a = 1, \quad \sigma = 2, \quad \gamma = 0.95.$$

This yields

$$\sqrt{n} \geq \frac{2 \cdot 2}{1} \Phi^{-1}\left(\frac{1+0.95}{2}\right) = 4 \cdot \Phi^{-1}(0.975)$$

$$= 4 \times 1.960 \quad \text{(in Z-table)}$$

$$= 4 \times 1.959964 \quad \text{(with R)}$$

Hence $n \geq (4 \times 1.960)^2 = 61.4656$

is a sufficiently large number of measurements

4.2 Sample Variance

If we make measurements of some quantity, we consider this as evaluating a RV X. If we make several measurements, then we consider them as evaluations of n RVs $X_1, \dots, X_n$ that are i.i.d., having the same distribution as X.

How can we estimate the mean value of the distribution of X, i.e., $E[X]$?

The average $\overline{X}_n$ of the X_i, $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$, should be a good estimate.

How can we check that this is conceptually the right thing to do?

Unbiased Estimators

Suppose $X, X_1, \dots, X_i, \dots$ are i.i.d. RVs.

A function $F(x_1, \dots, x_n)$, if applied to $X_1, \dots, X_n$, defines a new random variable $F(X_1, \dots, X_n)$.

An example is $\overline{X}_n$, which is defined by

$$F(x_1, \dots, x_n) = \frac{1}{n}(x_1 + \dots + x_n) = \overline{x}_n.$$

<u>Definition</u>: Let $X_1, \dots, X_n$ be i.i.d. RVs, $F: \mathbb{R}^n \to \mathbb{R}$ a function and θ be a parameter (like mean, variance, or skew) of the distribution of the X_i.

Then the bias of F with respect to θ for $X_1, \dots, X_n$ is

$$E(F(X_1, \dots, X_n)) - \theta,$$

and $F(X_1, \dots, X_n)$ is an unbiased estimator if the bias is 0.

Examples: (1) The average $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased estimator of the mean μ.

(2) The average squared distance from the mean

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

is an unbiased estimator of the variance. (Note that we used μ, not $\overline{X}_n$.)

Proof: (1) If have calculated several times that

$$E[\overline{X}_n] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\cdot n\cdot\mu = \mu$$

(2) Remember that $Var(X) = E[(X-\mu)^2]$. Thus

$$E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i-\mu)^2 = \frac{1}{n}\cdot n\cdot\sigma^2 = \sigma^2$$

Estimating the Variance

Consider the function

$$T^2(x_1,\dots,x_n) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \qquad {*)}$$

with $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Then one can calculate (see lecture notes of 19/20) that

$$E[\,T^2(X_1,\dots,X_n)] = \frac{n-1}{n}\,Var(X).$$

Thus, this is an estimator with bias! But

$$\frac{n}{n-1}T^2(x_1,\dots,x_n) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 =: S^2$$

is unbiased! This is also called the sample variance.

---

*) "T^2" is an abuse of notation, motivated by the attempt to estimate the variance

We can determine the quality of an estimate if we know how the random variable that we want to estimate is distributed.

Often, we assume that our X_i are $\mathcal{N}(\mu, \sigma^2)$-distributed.

In that case,

$$\overline{X}_n \quad \text{is} \quad \mathcal{N}(\mu, \tfrac{1}{n}\sigma^2)\text{-distributed.}$$

What about the distribution of

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \qquad ?$$

One can show that $\frac{n-1}{\sigma^2} S^2$ is distributed like the sum $Z_1^2 + \dots + Z_{n-1}^2$ of $n-1$ independent $\mathcal{N}(0,1)$ variables $Z_1, \dots, Z_{n-1}$. This is the χ^2_{n-1}-distribution.

The Interplay of Normal and Chi-squared

Theorem 67: Let $X_1, \dots, X_n$ be i.i.d $\mathcal{N}(\mu, \sigma^2)$. Then

- $\bar{X}$, S^2 are independent

  (sample mean: $\bar{X}$; sample variance: S^2)

- $\bar{X} \sim \mathcal{N}(\mu, \frac{1}{n}\sigma^2)$

- $\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$

The Chi-Square Distribution (χ^2-distribution)

The distribution of the sum of the squares $Z_1^2 + \dots + Z_n^2$ of n independent $N(0,1)$-RVs Z_i is called the Chi-square distribution of n degrees of freedom.

Notation: χ_n^2.

It almost follows from the definition that χ^2-distributions are reproductive:

$$X \sim \chi_m, \quad Y \sim \chi_n, \quad X, Y \text{ independent}$$

$$\Rightarrow \quad X + Y \sim \chi_{m+n}$$

There is a formula for the pdf of χ_n^2, but not the cdf

$\Rightarrow$ values have to be computed by numerical integration.

Mean and Variance of χ^2_n

Even without formulas for χ^2_n, we can find out the mean.

First consider χ^2_1. By definition, $Z^2 \sim \chi^2_1$.

- $Var(Z) = E[Z^2] + E[Z]^2 = E[Z^2] + 0^2$
- $Var(Z) = 1$

$$\Longrightarrow \quad E[\chi^2_1] = E[Z^2] = 1$$

Then

$$E[\chi^2_n] = E[Z_1^2 + \dots + Z_n^2] = E[Z_1^2] + \dots + E[Z_n^2] =$$

$$= n\,E[Z^2] = n \cdot 1.$$

Moreover, $\quad Var(\chi^2_n) = 2 \cdot n.$

What if we Don't know the Variance? t-Distribution!

We know that

$$X_i \sim N(\mu, \sigma^2) \implies \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \sim N(0,1)$$

What happens if we replace σ with $S = \sqrt{S^2}$?

$$\sqrt{n}\,\frac{\bar{X}-\mu}{S} = \frac{\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}}{\sqrt{\frac{1}{n-1}}\sqrt{\frac{(n-1)S^2}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$

A RV $T_n = \dfrac{Z}{\sqrt{\chi^2_n / n}}$ has a t-distribution with n degrees of freedom, written $T_n \sim t_n$

The t-Distribution: Definition

Suppose Z and χ^2_n are independent RVs and

- $Z \sim N(0,1)$
- $\chi^2_n \sim \chi^2_n$

Then the RV

$$T_n := \frac{Z}{\sqrt{\frac{\chi^2_n}{n}}}$$

is t-distributed with n degrees of freedom, written $T_n \sim t_n$

This uniquely defines cdf and pdf of the t-distributions.

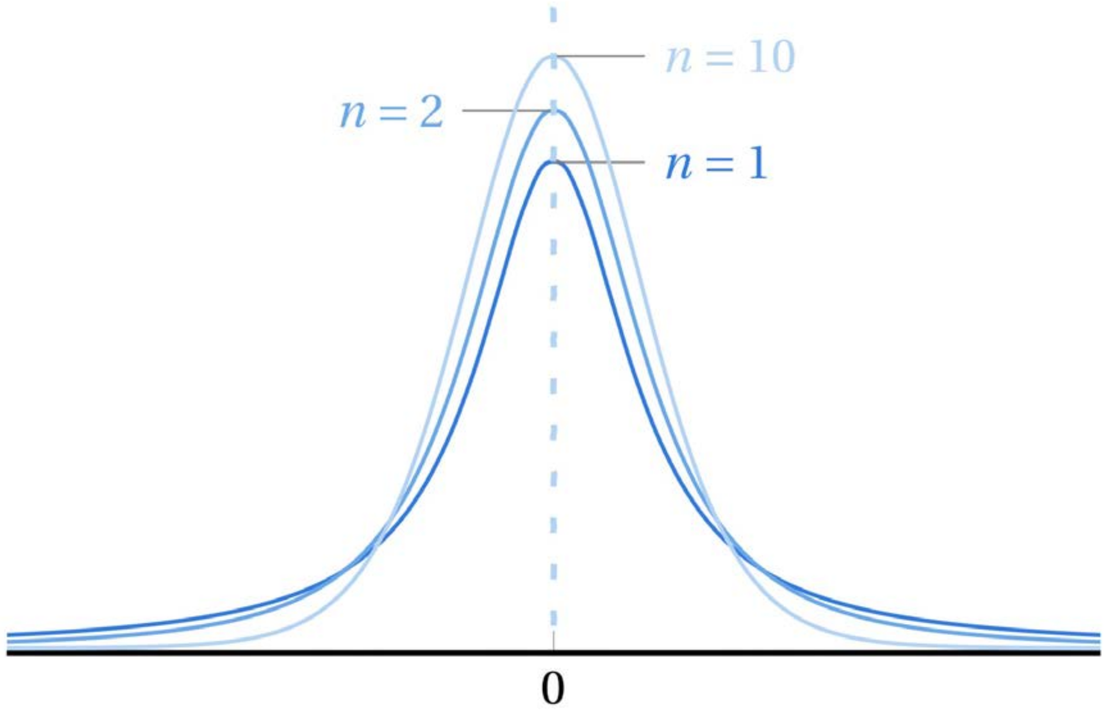
The t-Distribution: Properties

- Introduced by William Gosset (1908), chief brewer at Guinness, in a research paper published under the pen name Student. For that reason, it is also known as Student's t-distribution.

- It is bell shaped like the normal, but it is wider, the tail is thicker and the peak is lower than the peak of the standard normal. Its parameters are:

$$\text{mean} = \begin{cases} \text{undefined} & \text{for } n=1 \\ 0 & \text{for } n>1 \end{cases}$$

$$\text{variance} = \begin{cases} \text{undefined} & \text{for } n=1,2 \\ \dfrac{n}{n-2} & \text{for } n>2 \end{cases}$$



The density function of T_n for $n = 1, 2, 10$.

- It converges towards the standard normal, which can be seen as follows:

$$E[Z^2] = Var(Z) + E(Z)^2$$

$$= 1 + 0^2 = 1$$

If Z_i are i.i.d. $N(0,1)$, then $\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$.

By the law of large numbers, the average

$$\frac{\chi_n^2}{n} = \frac{1}{n} \sum_{i=1}^{n} Z_i^2 \longrightarrow E[Z^2] = 1 \quad (n \to \infty)$$

Therefore, also

$$\sqrt{\frac{\chi_n^2}{n}} \longrightarrow 1 \quad (n \to \infty).$$

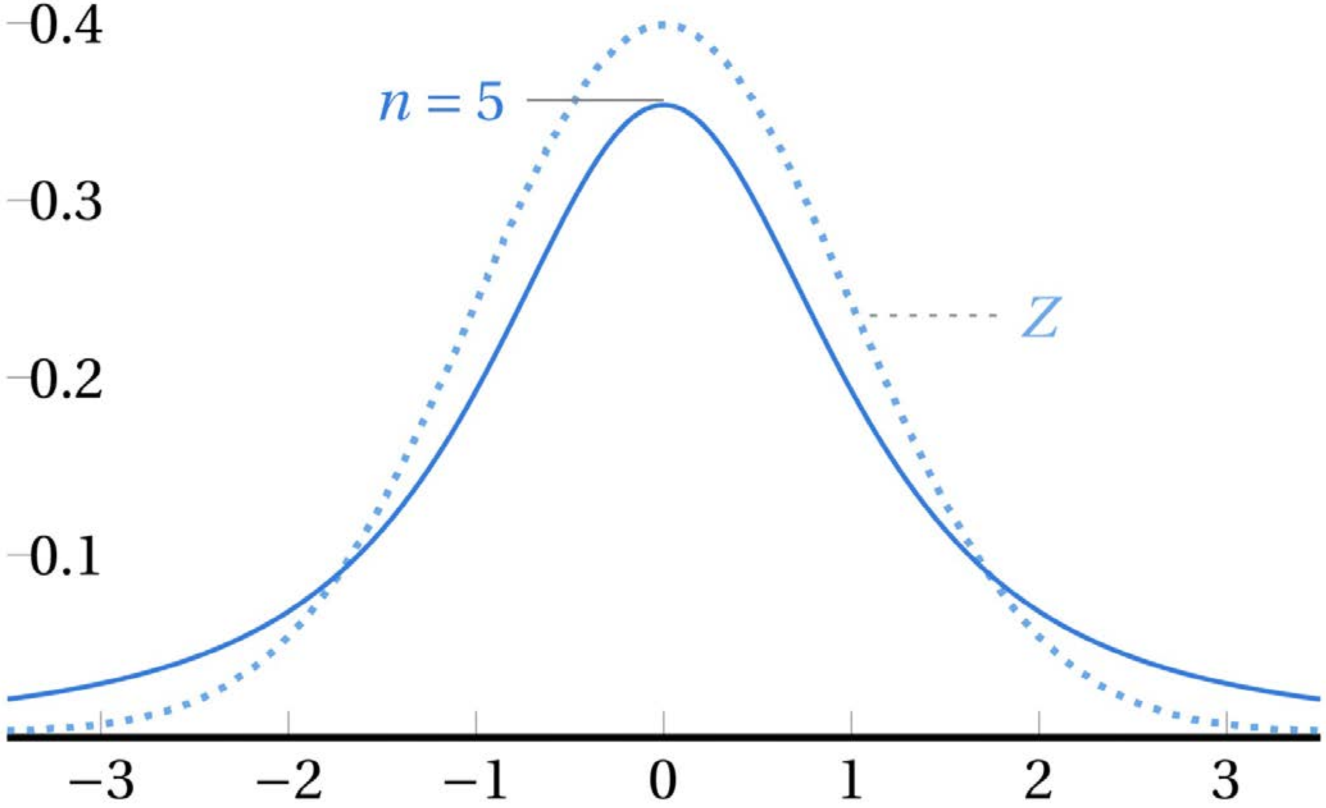
The convergence is "in probability", as expressed in the weak law of large numbers.

We now can apply Slutsky's theorem (see Wikipedia), since Z and $\sqrt{\frac{\chi_n^2}{n}}$ are independent:

$$\frac{Z}{\sqrt{\frac{\chi_n^2}{n}}} \longrightarrow \frac{Z}{1} = Z \quad (n \to \infty, \text{ in distribution})$$

The convergence is slower for the tail, that is, the farther we are away from 0. (Check the table of the t_n!)

- The t-distribution is practically relevant only for analyzing small samples (≤ 30), otherwise the difference to the normal is negligible.



The density function of T_5 (solid) and Z (dotted).